IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON AMBIENT INTELLIGENCE ENVIRONMENTS WITH WIRELESS SENSOR NETWORKS FROM THE POINT OF VIEW OF BIG DATA AND SMART & SUSTAINABLE CITIES

Received June 3, 2018, accepted July 7, 2018, date of publication July 19, 2018, date of current version August 15, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2857499

An Ontology-Oriented Architecture for Dealing With Heterogeneous Data Applied to Telemedicine Systems

JESÚS PERAL[1], ANTONIO FERRÁNDEZ[2], DAVID GIL[1], RAFAEL MUÑOZ-TEROL[1], AND HIGINIO MORA[3]

[1]Lucentia Research Group, Department of Software and Computing Systems, University of Alicante, 03690 Alicante, Spain
[2]Department of Software and Computing Systems, University of Alicante, 03690 Alicante, Spain
[3]Department of Computer Technology and Computation, University of Alicante, 03690 Alicante, Spain

Corresponding author: Jesús Peral (jperal@dlsi.ua.es)

This work was supported in part by the Spanish Ministry of Economy and Competitiveness (MINECO) under Project SEQUOIA-UA (TIN2015-63502-C3-3-R) and Project RESCATA (TIN2015-65100-R) and in part by the Spanish Research Agency (AEI) and the European Regional Development Fund (FEDER) under Project CloudDriver4Industry (TIN2017-89266-R).

ABSTRACT Current trends in medicine regarding issues of accessibility to and the quantity and quality of information and quality of service are very different compared to former decades. The current state requires new methods for addressing the challenge of dealing with enormous amounts of data present and growing on the Web and other heterogeneous data sources such as sensors and social networks and unstructured data, normally referred to as big data. Traditional approaches are not enough, at least on their own, although they were frequently used in hybrid architectures in the past. In this paper, we propose an architecture to process big data, including heterogeneous sources of information. We have defined an ontology-oriented architecture, where a core ontology has been used as a knowledge base and allows data integration of different heterogeneous sources. We have used natural language processing and artificial intelligence methods to process and mine data in the health sector to uncover the knowledge hidden in diverse data sources. Our approach has been applied to the field of personalized medicine (study, diagnosis, and treatment of diseases customized for each patient) and it has been used in a telemedicine system. A case study focused on diabetes is presented to prove the validity of the proposed model.

INDEX TERMS Ontology-oriented architecture, heterogeneous data, health sector, artificial intelligence methods, personalized medicine, telemedicine system, diabetes' treatment.

I. INTRODUCTION AND MOTIVATION

Healthcare has been generally highly connected to technology. However, this relationship has become stronger over the last two decades. One of the main reasons for this is the proliferation of all types of devices that can be easily installed in most health centers. In addition, telemedicine, which was first mentioned several decades ago, is now a reality and has been highly developed and this evolution has also extended to other healthcare sectors. Moreover, in recent years mobile devices are able to obtain a number of biomedical data and health-related apps have increasingly been developed [1]–[3]. In addition to smartphones, there are a lot of gadgets that use sensors to collect information from different parts of the human body. In this respect, the paper ''Infographic: are you ready for sensors healthcare'' provides a novel approach to the distribution of sensors around the body so as to gather the most accurate information [4]. The market for and advertising of wearable sensors is rapidly developing and growing, as studied in the emerging Internet of thing, IoT, where a huge amount of data can be collected everywhere, all the time [5]–[7], which requires to store the heterogeneous information generated in cloud storage solutions as the ones proposed in [8]–[10]. Devices are being designed for the following purposes: to help people manage particularly chronic conditions; to recover faster from injuries, a well-researched sector in the sports market; to analyse physical and environmental anomalies that may lead to more serious health issues; and to detect unhealthy habits before they cause problems, an aspect is taken very seriously in the analysis of working conditions according to Pathfinder [11], even with those

issues related to energy efficiency for the ubiquitous sensors networks [12].

New trends in cognitive science are still a challenge as they require a huge degree of interdisciplinarity. With changes to models of society and new technologies used in telemedicine as well as in all healthcare sectors, the traditional methods seem insufficient for this new situation. This new state requires using new techniques to address the problem. Although, traditional methods are still valuable, they need to be used together, providing new architecture models.

The main objective of this paper is to assemble a composite architecture that integrates the diverse information sources currently available in the healthcare sector. The proposed architecture focuses on the processing of the enormous amounts of data present and growing on the Web and other heterogeneous data sources such as sensors and social networks and unstructured data, normally referred to as Big Data. We have used an ontology-oriented architecture, where the core ontology has been defined as a knowledge base (KB), which allows data integration of the different sources. We have described the data sources in terms of the core ontology through equivalent concepts and relations (ontology mapping) between core ontology and specialized domain ontologies of the different sources. We propose the use of Artificial Intelligence (AI) methods to carry out a Data Mining (DM) process for the purpose of uncovering knowledge in the healthcare sector.

Our proposal has been applied to the field of personalized medicine, in which the treatment of diseases is customized for each patient, and it has been used in a telemedicine system. Finally, we present a case study based on the treatment of diabetes to demonstrate the validity of the proposed model.

The contributions of our paper can be summarized as follows:

- An architecture that processes massive amounts of data from diverse types of sources (heterogeneous data sources). Furthermore, it allows the semantic-level integration of heterogeneous sources of information.
- An ontology-oriented architecture has been defined where the core ontology has been used as a KB enabling data integration of the different sources.
- Natural Language Processing and Artificial Intelligence methods have been used in order to process and mining data in the healthcare sector to uncover knowledge from diverse data sources.
- Application of the approach to the field of personalized medicine (study, diagnosis, and treatment of diseases customized for each patient). An existing real healthcare problem was solved.
- The proposed model was validated through a case study focusing on diabetes.
- A telemedicine system that helps physicians in the decision-making process in the treatment of diabetes has been presented in this case study. The system allows the physician to improve the rules for treatment procedures learned using AI techniques.
- The use of different sources of information (information stored in databases, Web information and sensor information) allows for improvements (oriented towards patient personalization) in the rules to be applied for each specific patient.

The main novelties presented are:

- An ontology-oriented architecture that uses a central ontology that permits communication between different data sources each with its own ontology.
- An improvement of the traditional AI systems on diabetes' treatment. The personalized treatment of each patient and the improvement of traditional AI systems have been made possible by including different data sources.
- The application of the proposed architecture in the telemedicine system in order to improve its performance.

The remaining part of the paper is organized as follows: in Section II, the health-related scientific literature is summarized; in Section III, we present our approach, that is, the proposed architecture; Section IV describes the application of our proposal to the health sector case study; finally, Section V draws the relevant conclusions and presents future work.

II. STATE OF THE ART

As introduced in the preceding motivation section, we present an architecture that integrates different heterogeneous sources currently available in the health sector. In this regard, as presented in the paper by De Buenaga Rodríguez *et al.* [13], the latest advances and discoveries in the biomedical field, both in terms of technology and basic research, have led to important progress in the approach to, and modern clinical practices of, Evidence-Based Medicine, EBM [14] and personalized medicine. Personalized medicine seeks to identify personalized therapies to provide a safe and effective individualized treatment of specific patients. To allow for personalized treatment, it is first necessary to make a correct personalized diagnosis. Currently, this seems to be the most suitable framework due to a large amount of information available (experimental studies, clinical trials, daily clinical practice, biomedical sensors, large datasets and text freely available–open and Linked Data–, etc.). However, the real situation is that there are no flexible information systems capable of providing accurate, updated and interrelated knowledge based on stratified access to multiple types of heterogeneous data sources [15]. Proper management of this extraordinary source of knowledge would provide a breakthrough in the correct diagnosis and personalized treatment. For example, the research project currently underway and reported in [13] foresees three usage scenarios: (i) assistance to healthcare professionals during the decision-making process in clinical settings, (ii) access to relevant information about their health status and dependent chronic patients and (iii) support for evidence-based training of new medical students. They propose the integration of biomedical sensors as detailed in [16]. Moreover, that is not

all because different research efforts have been developed last years in the wide range of health sector. We can mention some examples of these ones such as the works developed by Medina *et al.* [17], [18]. They present the application of a fuzzy linguistic approach over medical monitoring devices on data streams in the development frame of a multi-dose medication controller for fever. Their approach defines fuzzy linguistic terms on single medical monitoring devices and applies a Rule-Based Inference Engine designed for analysing the data streams. On the other hand, the solution exhibited by Urzaiz *et al.* [19] details the combination of hardware and software environments to develop an automatic medication dispenser for patients with Alzheimer's disease.

We have divided the diverse information sources into different labels which are: (1) unstructured data (Web data) and the techniques used to process them (Natural Language Processing); (2) structured information (structured databases) and the processes carried out to extract information from them (Data Mining); and (3) data collected from sensors and mobile devices (sensor data). It is worth noting that although the sensor data could be classified as structured databases, we have decided to analyse it separately (as a new type of data source) because of its own specific characteristics.

The following three subsections discuss the state-of-the-art of the mentioned issues regarding the health sector: Web data, structured databases and Data Mining, and sensor data. Finally, the fourth subsection shows the state-of-the-art of the problem of data integration.

A. WEB DATA AND NATURAL LANGUAGE PROCESSING

Nowadays, different Natural Language Processing (NLP) systems and techniques are being used in biomedical and health sectors. Specifically, Savova *et al.* [20] presented a clinical text analysis and knowledge extraction system for extracting information from electronic medical records' free-text. In the named entities recognition task, a dictionary that is a subset of the Unified Medical Language System, UMLS [21] is used to include SNOMED CT[1] and RxNORM[2] concepts guided by extensive consultations with clinical researchers and practitioners. In the same way, Pivovarov *et al.* [22] present a probabilistic graphical model for large-scale discovery of computational models of disease, or phenotypes in which the observations are drawn directly from heterogeneous patient record data (notes, laboratory tests, medications, and diagnosis codes) and Soguero-Ruiz *et al.* [23] developed a learning system which uses clinical narrative in free text form for the prediction of a common postoperative complication (Anastomosis Leakage).

The research work developed by Zeng *et al.* [24] presented a health information text extraction tool that is used to extract key findings for a research study on diseases of the airways. This NLP tool also maps the strings of text to the UMLS concepts. UMLS [25] is a resource that integrates and distributes key terminology, classification and coding standards, and associated resources to promote the creation of more effective and interoperable biomedical information systems and services, including electronic health records. UMLS contains three knowledge sources: the Metathesaurus, the Semantic Network, and the Specialist Lexicon. In particular, the UMLS Metathesaurus [26] is used to recognize medical named entities in the text using a similar process to the one presented by Terol *et al.* [27].

Wu *et al.* [28] showed an NLP technique that characterizes empirical instances of the UMLS Metathesaurus term strings in a large clinical corpus and illustrates what types of term characteristics are generalizable across data sources. The study developed by Xu *et al.* [29] analysed the UMLS Metathesaurus terms by analysing their occurrences in over 18 million MEDLINE abstracts. This study concluded with an augmented UMLS Metathesaurus that can potentially be used to improve efficiency and precision of UMLS-based information retrieval and NLP tasks. Jiang *et al.* [30] developed different machine-learning-based approaches to extract clinical entities (including medical problems, tests, and treatments, as well as their asserted status) from hospital discharge summaries written using natural language. The studies developed by Carroll *et al.* [31] concluded with the portability of a published phenotype algorithm to identify rheumatoid arthritis (RA) patients from Electronic Health Records (EHR) at three institutions with different EHR systems. They concluded that generic UMLS NLP systems may be sufficient for good performance in at least some specific phenotype identification tasks. Another interesting work is the one in [32], which automatically extracts explicit knowledge from databases under the form of IF-THEN rules containing AND-connected clauses. This knowledge is applied for diagnosis in the medical domain.

Thus, by analysing all these NLP systems and techniques, we can conclude that most of them use UMLS as a knowledge source in the biomedical and health domains.

B. STRUCTURED DATABASES AND DATA MINING

Structured databases allow DM and Machine Learning (ML) techniques to be applied directly for extracting the relevant knowledge. In short, DM is the process of analysing data from various perspectives and summarizing it into useful information [33]. The hidden information that is made available through data mining can benefit the person involved by providing efficient decision support. It is stated that providing decision support in the healthcare sector can help save human lives [34]. Data-driven healthcare opens up new opportunities in personalised medicine, preventive care, chronic disease management and in telemonitoring and managing patients with implanted devices [35].

In addition, extracting knowledge from information and data has been the main goal in a lot of work related to data management dealing with several different application areas including healthcare [36]. Machine learning tech-

[1]https://www.nlm.nih.gov/healthit/snomedct/index.html (visited on 19th of March, 2018).

[2]http://www.nlm.nih.gov/research/umls/rxnorm/ (visited on 19th of March, 2018).

niques are being intensively applied to the healthcare sector for predictive analysis. These techniques are classified into three categories: (1) association rule, (2) classification and (3) clustering [37]. A survey on DM approaches for healthcare can be found in [38].

Association rules refer to the discovery of relationships between elements. For example, they may discover that a set of indications or symptoms frequently occur together with another set of symptoms. A priori is a typical method in this category. Some recent examples of the application of association rule DM to different sectors of healthcare can be found in [33], [39], and [40].

Classification maps data items into one of several predefined classes. For example, classification rules about a disease can be extracted from previous known cases and then used to diagnose new patients of the disease based on their symptoms. Decision Trees [41], [42], Artificial Neural Networks [42], [43], Support Vector Machines [**?**], [44], [45], Bayesian Networks [46] and Naive Bayes [47] are examples of classification approaches applied to healthcare.

Clustering recognizes the class for a set of unclassified elements according to their attributes. For example, a set of diseases can be grouped into several clusters based on the similarities in their symptoms, and the common symptoms of the diseases in a cluster can be used to describe or predict that group of diseases. K-nearest neighbour (K-nn) is one of the most popular methods of clustering. Examples of application in healthcare can be found at [48] or [49].

Other artificial intelligence techniques include evolutionary methods [50]. For example, in [51], a genetic algorithm is used as part of the process for identification of patient phenotype cohorts, by mining textual data from health records.

Another important set of new techniques comes from Web data mining. Web data mining aims to uncover useful information and knowledge from the web hyperlink structure, page contents, and usage data [52]. One recent example of its application for disease prediction can be found in [53], where a remodeling of HITs algorithm is proposed.

Datasets used in many current health informatics studies can be categorized as Big Data. Big Data can be defined as including the following qualities: Volume, Velocity, Variety, Veracity, and Value [54]. Volume refers to the large size of datasets, Velocity refers to the great speed with which new data is incorporated, Variety refers to the different formats and structures used for data representation, Veracity refers to correspondence with reality, and Value refers to quality of the dataset in correlation to the intended result.

In the healthcare sector, there are an increasing relevant big datasets coming from, among others: widespread uptake of electronic health records (EHR) [34], [55]; clinical sensors, some of them wearable, used in new health monitoring systems [56]–[58]; health relevant environmental data, such as pollution, tobacco smoke, pollen, now available through mobile devices and wireless networks [59]; social networks and online services [60], [61].

One problem in relation to the mentioned datasets is the design of architectures and frameworks allowing record processing at a reasonable throughput in a scalable way. Generic big data solutions such as the MapReduce framework, distributed file systems or NoSQL databases have been successfully applied to some health informatics problems in diverse proposals ([62]–[64]). In [65], these technologies and their possibilities in the clinical data analysis are described.

We conclude this section by referring to an additional problem caused by the general availability of a plethora of databases, resulting in the need to integrate or merge different databases [66], [67]. Merging poses several problems. For example, duplicated tuples need to be removed and replaced with a single tuple that represents the joint information of the duplicate tuples to the greatest extent ([68]), which can breach referential integrity. This fusion also poses additional problems such as typographical errors, lack of standardization and missing data, making it a considerable task. The problem is even greater when heterogeneous data sources are merged [69]. For example, Bilke *et al.* [70] and Naumann *et al.* [71] proposed the HumMer system for the semi-automatic integration of heterogeneous data sources. It distinguishes three steps in the process of data integration: schema matching, duplicate detection, and data fusion. In [72] additional merge problems are described: when the database models are different (e.g. relational database vs. object-oriented database), different data schemas (i.e. different ways of describing data or different codification criteria), mismatched domains, semantic conflicts, semantic reconciliation, schema integration, etc.

C. SENSOR DATA

Recently, different approaches have been developed in the communication between biomedical sensors and smartphones. The system architecture developed by Cobelli *et al.* [73] presents two means of communication in the frame of Type 1 Diabetes: the first is communication between the Continuous Glucose Monitor (CGM) sensor and the smartphone, and the second is communication between the smartphone and the insulin pump. The proposal by [74] faces the critical characteristics of Diabetes management as each diabetic is a unique case with specific needs. Through data mining algorithms their proposal provides individual advice to the diabetic user. An interesting contribution of their work is that when the number of records is not enough to attain useful conclusions, their proposal uses a set of logical rules, defined from medical protocols, directives. Moreover, in the frame of cholesterol measurements, Oncescu *et al.* [75] present a system that can be used to measure and track cholesterol levels directly on a smartphone throughout the optimal image acquisition and processing. Also, Yi *et al.* [76] show a communication architecture between the smartphone and different sensors such as a temperature sensor and an electrocardiography (ECG). Other approaches in the communication scenario

between smartphones and biomedical sensors is presented by Fortino and Giampa [77] who present the Continuous Blood Pressure Measurement. Finally, mention the paper by [78] that presents a multi-sensors system to enhance the accuracy of potentially dangerous heart rate variability by considering patient context information.

With the aim of obtaining a data set of human activity recognition, the previous research work developed by Reyes-Ortiz *et al.* [79] presents the set of experiments that were carried out with a group of 30 volunteers within an age bracket of 19-48 years. The group performed a protocol of activities composed of six basic activities: three static postures (standing, sitting, lying) and three dynamic activities (walking, walking downstairs and walking upstairs). The experiment also included postural transitions that occurred between the static postures. These are: stand-to-sit, sit-to-stand, sit-to-lie, lie-to-sit, stand-to-lie, and lie-to-stand. All the participants were wearing a smartphone (Samsung Galaxy S II) on the waist during the experiment execution. They captured 3-axial linear acceleration and 3-axial angular velocity at a constant rate of 50Hz using the embedded accelerometer and gyroscope of the device. The sensor signals (accelerometer and gyroscope) were preprocessed by applying noise filters and then sampled in fixed-width sliding windows of 2.56 sec and 50% overlap (128 readings/window). The sensor acceleration signal, which has gravitational and body motion components, was separated using a Butterworth low-pass filter into body acceleration and gravity. The gravitational force is assumed to have only low frequency components, therefore a filter with 0.3 Hz cutoff frequency was used. From each window, a vector of 561 features was obtained by calculating variables from the time and frequency domain.

The experiments were video-recorded to label the data manually. The obtained dataset was randomly partitioned into two sets, where 70% of the volunteers were selected for generating the training data and 30% the test data.

Moreover, smartphones have also been used to measure physical activity and the sedentary lifestyle levels of people. Regular physical activity is known to help prevent and treat numerous non-communicable diseases like diabetes. Smartphone apps have been shown to increase physical activity in primary care. It is a fact that nowadays there are many apps that can report on the degree of physical activity of users. For example, variables like speed walking, running, climbing stairs and physical activity duration can be measured by new healthy lifestyle apps.

D. DATA INTEGRATION

As presented in [80] and [81], in the past many techniques have been developed on data integration of different heterogeneous data sources:

- Most research on data integration has focused on the relational model. In many ways, the relational model and the datalog query language are the simplest and cleanest formalisms for data and query representation, so many of the fundamental issues were considered first in that setting. [82]–[87].
- XML (W3C recommendations)[3] has become the default format for data export from both database and document sources, and many additional tools have been developed to export to XML from legacy sources. However, XML is not meant to directly resolve semantic heterogeneity or introduce standard schemas in any domain [88]–[93].
- Uncertainty can be introduced in multiple aspects of data integration: (a) Data, some of the data may be extracted from unstructured data and we may be uncertain about the accuracy of extraction; (b) Schema mappings, may be generated using semiautomatic techniques and we may not have the resources to validate all these mappings. Uncertainty about schema mappings can be very common in data integration applications; (c) Queries, the system may need to offer the user a keyword-query search interface and needs to translate the keyword queries into some structured form, so they can be reformulated onto the data sources. The translation step may generate multiple candidate structured queries, and therefore there will be uncertainty about the intended user query; (d) Mediated schema, when the domain of the integration system is very broad, there may even be some uncertainty about the mediated schema itself. [94]–[99].
- The Knowledge Representation (KR) system stores the underlying model of the domain that is employed by many artificial intelligence applications, such as planners, robots, natural language processors, and game-playing systems. KR systems use reasoning techniques to answer queries about knowledge. Some aspects of data integration can also be viewed as a knowledge representation problem. Data sources and their contents lend themselves to rather complex modeling. Determining the relationships between data sources, or between a data source and a mediated schema, often requires subtle reasoning. For these reasons researchers have considered applying knowledge representation techniques to data integration. Description logics is the main family of KR languages that has been employed in the context of data integration and it offers a set of logical formalisms for defining a domain ontology (a description of the entities in the domain, relationships between them, and any other known constraints). [100]–[104]. Finally, the work developed by Ganzha *et al.* [105] performs a rich study about available and ready to be used ontologies for the development of interoperable applications in the IoT environment by the interrelationships across the different heterogeneous data sources.

[3] http://www.w3.org/ (visited on 19th of March, 2018).

III. OUR APPROACH. ARCHITECTURE PROPOSAL

The preceding section reviews scientific literature regarding the different information sources and their integration. Due to the heterogeneity of these datasets, we believe that in this case, it is necessary to address the problem with new methods and approaches. New techniques are fundamental as the now traditional methods are more inadequate and restricted mainly due to the current big data situation. One of the main characteristics and advantages of our proposed architecture is the use of NLP and DM techniques that allow the processing, the integration and the extraction of relevant information of the different sources.

The architecture proposed in this paper is summarized in Figure 1, which shows a diagram representing the methodology employed in this study.

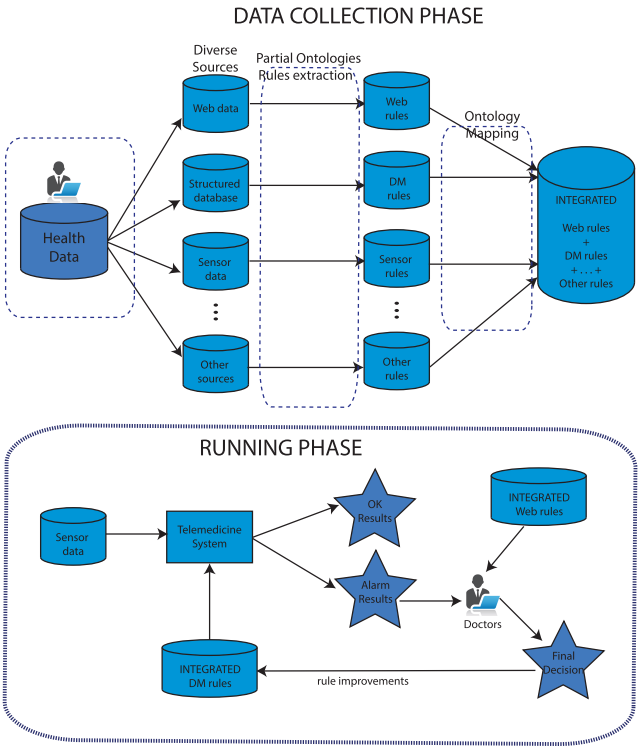

FIGURE 1. General proposed architecture. Data collection and running phases.

We have used an ontology-oriented architecture. The core ontology has been defined as a KB and allows data integration of the different sources. This integration will be made identifying the equivalent concepts and relations (ontology mapping) of the different sources. In our methodology we can distinguish two phases: (1) data collection phase and (2) running phase.

In the data collection phase the information from various data sources is collected. Specialized domain ontologies for each specific data source are used. A process, called rules extraction, is applied to each different source. In the case of unstructured information (Web data) NLP techniques are used for the purpose of obtaining the Web rules. Regarding to structured information (structured databases) NLP and DM techniques are used to obtain DM rules. Finally, all the information is integrated with the process of ontology mapping in which the equivalences between the concepts of the data sources are established. The result of the data collection phase is a set of different kinds of rules that we have called integrated rules (integrated Web rules, integrated DM rules, etc.). All these rules are specified using the same concepts, that is, the concepts of the core ontology.

In the running phase the information collected from the sensors (sensor data) and the integrated DM rules are sent to the telemedicine system. If the system detects any abnormal measurement it sends the corresponding alarm to the medical team. The physician analyses the alarm by consulting the integrated Web rules. Finally, they will act: (1) rejecting the alarm because it is an exception, or (2) accepting the alarm. All this information will be considered in order to improve and refine the integrated DM learned rules.

In the next two subsections the abovementioned processes of rules extraction and ontology mapping are described in detail.

A. THE RULES EXTRACTION PROCESS

We can distinguish different extraction processes depending on the information source. In this paper we focus on unstructured and structured data. Therefore, the process to extract the rules from Web data and from structured databases are defined below.

1) WEB RULES EXTRACTION

In order to manage the great quantity of unstructured information processed by the system and to help physicians in their decision-making process, the NLP module performs a set of tasks with the aim of giving physicians sufficient information in terms of quantity and quality needed to make the right decisions in the treatment of patients. These tasks combine different NLP techniques as described below:

- Document selection. With regard to the information extracted from the Web, documents related to the health sector are selected.[4]
- Information Retrieval (IR) Task. This IR task consists in finding those documents that satisfy the physicians' information needs from the large document collections. Our approach uses the term frequency-inverse document frequency (tf-idf) technique to accomplish its goal. In this technique, each word has a statistical measure (weight) assigned to it, which is used to evaluate how important a word is to a document collection. Importance increases proportionally to the number of times a word appears in the document but is offset by

[4]For instance, the following URLs' corpus could be crawled (visited on 19th of March, 2018): PubMed (https://www.ncbi.nlm.nih.gov/pubmed/); Artificial Intelligence in Medicine (https://www.journals.elsevier.com/artificial-intelligence-in-medicine/); PLOS ONE (http://journals.plos.org/plosone/); Journal of Biomedical Informatics (http://www.journals.elsevier.com/journal-of-biomedical-informatics/).

the frequency of the word in the collection. Therefore, according to the information needs of physicians formulated as an utterance in natural language, the IR task retrieves those documents that contain the most number of words from the given utterance in the tf-idf frame explained. This task allows huge volumes of text to be filtered with the aim of decreasing the quantity of text in which computational expensive NLP techniques will be applied.

- Ontology-Based Semantic tagging and Retrieval Tasks. After filtering the document collections, the information request posed by physicians is semantically analysed with the aim of tagging the possible semantic types expected in the documents. On the one hand, this task consists of using the domain ontology as a knowledge source to label words and utterances contained in the documents retrieved in the previous IR task.[5] On the other hand, once the words and utterances of the summaries have been semantically labelled as entities, the task ranks these documents according to the amount of semantic types expected in the documents. Thus, based on this ranking, the system assumes that the top documents are candidates for containing sufficient information in terms of quantity and quality that physicians need for their decision-making process.
- Information Extraction Task. This task performs a specialized Information Extraction process in the medical domain. Previously, the selected documents were partial syntactic parsed –only noun phrases (*np*), verbal phrases (*vp*), and prepositional phrases (*pp*) are identified–. Furthermore, the named entity recognition process was carried out in order to tag the noun phrases –for instance, GA (glycated albumin) and HbA1c are tagged as ''protein'' entities whereas DPN (diabetic peripheral neuropathy) is tagged as ''disease'' entity–.
- Ontology mapping. The different data sources will have their own ontologies or semantic resources, referred to above as domain ontologies. As previously mentioned, the information of each source will be semantically enriched/associated with the information extracted from their specific domain ontology. Furthermore, the core ontology has been defined as a KB. This ontology will be used when the integration of all information is carried out. To accomplish this integration, data sources have been described in terms of the core ontology through the equivalent concepts and relations (ontology mapping) between the core ontology and the specialized domain ontologies of the different sources. This process will be explained in detail in the next subsection.
- Extraction of the integrated Web rules. Finally, the integrated Web rules that use core concepts are extracted thanks to a series of predefined patterns described in terms of the core ontology concepts. The rules contain relevant information about the doctors' needs. This result is provided to the specialist physicians. With this information, they can create a personalized treatment for the patient depending on their personal characteristics with a more accurate analysis of the alarms produced by the telemedicine system. The Web rules will be later used by physicians and computer experts to improve the DM rules used by the telemedicine system.

2) DM RULES EXTRACTION

With the aim of gathering the DM rules from structured information several NLP and DM tasks are carried out:

- Document selection. With respect to structured databases, databases related to the health sector could be used.[6]
- Ontology-Based Semantic tagging. Once the database has been selected, the objective of this task is to identify the semantic concept of each field of the database in order to subsequently perform the integration of different data sources. All the database attributes are searched in the domain ontology obtaining their semantic concepts. As previously mentioned, WSD techniques should be applied if it is necessary to obtain a unique semantic concept.
- Data Mining task. As already mentioned in Section II-B, there are three main Machine Learning techniques. From these, we have decided to use classification techniques because they are the most adequate strategies for solving the problem of diagnosis and treatment of diseases based on patients' symptoms (our case study). The results of this task for structured databases is the classification of data into pre-defined classes. Depending on the specific problem, we can use different classification approaches cited in Section II-B. In the experimentation of this paper, we have used decision trees to obtain a set of rules that determine the diagnosis or treatment of a patient's disease.
- Ontology mapping and acquirement of the integrated DM rules. In this task, the rules obtained after the DM process are expressed in core concepts after the ontology mapping. These rules are called integrated DM rules and they are used by the telemedicine system with the objective to check the information collected by the sensors (sensor data) detecting any abnormal measurement.

B. THE ONTOLOGY MAPPING PROCESS. DATA INTEGRATION

The ontology can represent complex relationships between data sources, the mediated schema, and queries. The inference engine can reason these relationships to decide whether

[5] If a concept had different semantic types, Word Sense Disambiguation (WSD, [106], [107]) techniques should be applied to obtain a unique semantic type.

[6] For example, the following URLs' repositories could be processed (visited on 19th of March, 2018): UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/index.php); EU Open Data Portal (https://data.europa.eu/euodp/en/data/dataset?q=diabetes &ext_boolean=all); data.world open portal (https://data.world/).

data sources are relevant to given queries. Furthermore, ontologies can resolve semantic heterogeneity (unlike XML format). For these reasons, ontologies have been chosen as an integrator mechanism of all data sources in our proposal and, consequently, we present an ontology-oriented architecture.

The ontology will serve as the mediated schema of the data integration system. The different data sources will have their own ontologies or semantic resources, referred to above as domain ontologies. Thus, the information of each source will be semantically enriched/associated with the information extracted from their specific domain ontology.

Furthermore, the core ontology has been defined as a KB. This ontology will be used when the integration of all information is carried out. To accomplish this integration, data sources have been described in terms of the core ontology through the equivalent concepts and relations (ontology mapping) between the core ontology and the specialized domain ontologies of the different sources.

With the objective to select the core ontology, we have based on the concept of Universal Ontology (UO) presented by Olivé [108]. By UO Olivé means the formal specification of all the concepts that we use and share. This includes the concepts of general use, those that are particular to the existing disciplines, and those specific to any kind of human or organizational activity. The UO specifies the concepts that apply to objects, to their relationships, and to the actions or events involving those objects. In the UO four levels of concepts are distinguished: (1) Conceptual Model, (2) Foundational, (3) General, and (4) Domain. If we arrange vertically the levels and populate each level, the result can be interpreted as a pyramid, called the UO pyramid, in which the base is the Domain Level. Each of these levels is briefly described below:

1) The Conceptual Model (or Ontology Model) Level comprises the meta types and the direct or indirect supertypes of all the concepts in the UO. The concepts at this level are used to define the rest of the UO.
2) The Foundational Level includes abstract concepts that have been proposed in the foundational ontologies. The concepts at this level cannot be directly instantiated to publish facts and, therefore, they are not essential in the proposed UO. However, they may be useful for clarifying the semantics of other concepts, for defining only once knowledge that is common to several concepts, and for reasoning purposes.
3) The General Level contains the concepts for general purposes. These concepts are subtypes or instance of concepts at the conceptual Model Level, and, possibly, of concepts at the Foundational Level. There are several ontologies that could provide an excellent basis from which to build the General Level of the UO. Among them, we could mention WordNet [109] and CYC [110].
4) The Domain Level contains the concepts corresponding to the languages for special purposes. Therefore, this level contains all existing domain ontologies. Since there are many domain ontologies the Domain Level includes several millions of concepts. Achieving a satisfactory arrangement of these ontologies is the main technical challenge of the UO. The concepts at the Domain Level are subtypes or instance of concepts at the General Level, and, possibly, of concepts at the Foundational Level.

According to the UO presented, in our proposal we have used WordNet as the core ontology. It is a perfect basis to create the General Level of the UO. WordNet defines noun, verb and adjective synsets that may be the source of the entity types and properties of the UO. WordNet 3.0 comprises over 80,000 noun synsets (concepts), which include most entity types that have a name in English (general purpose). There are already ''wordnets'' in many languages, which include links to the English WordNet. It comprises also over 13,000 verb synsets, which include most properties that have a name in verb form. Finally, WordNet comprises also over 18,000 adjective synsets, most of which can be considered as Boolean properties.

Concerning the ontology mapping process between the specialized domain ontologies and the core ontology, there are two kinds of mappings: vertical and horizontal [108]. The vertical mappings define the correspondences between the domain ontology and the concepts at the general level (WordNet in our approach). The horizontal mappings define the correspondences between the domain ontology and the other ontologies at the domain level. In both mappings, a correspondence is a relationship between two concepts. In general, it can be an *equivalence* (the concepts are the same), an *IsA* (a concept is a subtype of the other) or a *disjointness* (no entity–or property– can be an instance of both concepts) [111].

In our approach, we propose the use of STROMA methodology [112] (SemanTic Refinement of Ontology MAppings) to determine both vertical and horizontal automatic semantic ontology mappings. The authors present a two-step methodology that leverages the capabilities of state-of-the art ontology match tools. In a first step, they apply a state-of-the-art match tool to determine an initial ontology mapping with approximate equality correspondences. In the second step authors apply five different techniques (including linguistic approaches and the use of dictionaries) to determine for each correspondence its most likely kind of relationship (*equality*, *is-a* –subsumption–, *inverse is-a*, *part-of* and *inverse part-of* relations). The five implemented strategies are the following: (1) Compound strategy that processes the compound words (head + modifier); (2) Background Knowledge strategy which uses linguistic resources or dictionaries (WordNet, OpenThesaurus[7] and parts of the UMLS Metathesaurus)[8] to check the relations; (3) Itemization strategy that is used if at least one of the two concepts in a correspondence is

[7]https://www.openthesaurus.de/ (visited on 19th of March, 2018).

[8]https://www.nlm.nih.gov/research/umls/ (visited on 19th of March, 2018).

an itemization (a list of items); (4) Structure strategy which takes the explicit structure of the ontologies into account; (5) Multiple Linkage is a specific strategy that draws conclusions of schema elements participating in more than one correspondence.

IV. CASE STUDY

A. OVERVIEW

Now that we have introduced the system architecture (Figure 1), in the following subsections we explain the application of our architecture to the healthcare environment scenario. Specifically, we focused on a telemedicine system to facilitate the diagnosis and treatment of patients with diabetes.

First of all, the context of the disease (Section IV-B) is presented. Next, Section IV-C describes the dataset used in our experiments. Following this, the data collection phase is explained, explicitly the rules extraction process: Web rules extraction (Section IV-D) and DM rules extraction (Sections IV-E). Finally, Section IV-F presents the running phase which includes sensor data collection and how the telemedicine system works.

B. THE CONTEXT

In our case study we will focus on Diabetes Mellitus (DM), commonly referred to as diabetes. It is a group of metabolic diseases in which there are high blood sugar levels over a prolonged period. Symptoms of high blood sugar include frequent urination, increased thirst, and increased hunger. If left untreated, diabetes can cause many complications. Acute complications include diabetic ketoacidosis and nonketotic hyperosmolar coma and are the consequence of an inadequate control of the disease. Serious long-term complications include cardiovascular disease, stroke, chronic kidney failure, foot ulcers, and damage to the eyes.

The major long-term complications of diabetes relate to damage to blood vessels. The primary complications of diabetes due to damage in small blood vessels include damage to the eyes (diabetic retinopathy), kidneys (diabetic nephropathy), and nerves (diabetic neuropathy). Regarding ''macrovascular'' diseases, complications include stroke, and peripheral vascular disease.

There are three main types of Diabetes Mellitus: (1) Type 1 DM (''insulin-dependent diabetes mellitus''–IDDM–) results from the pancreas' failure to produce enough insulin; (2) Type 2 DM (''non insulin-dependent diabetes mellitus''–NIDDM–) begins with insulin resistance and it produces a progressive defect in insulin secretion; (3) Gestational diabetes occurs when pregnant women without a previous history of diabetes develop a high blood-sugar level.

In our experiment we will study patients with Type 1 Diabetes. A telemedicine system will be used to track the disease and the treatment of diabetic patients with the objective of preventing the abovementioned complications that are typical of this disease from occurring. Our efforts will be aimed at achieving personalized medicine for each patient. Regarding the personalized treatment of diseases three possible scenarios can be identified: (1) assisting health professionals during the decision-making process in clinical environments; (2) providing chronic and dependent patients with access to relevant information about their health; (3) supporting evidence-based training for new medical students.

From these scenarios we will focus on the following:

1) Physician's perspective: the treatment that professionals give a diabetic patient. The goal of diabetes treatment is to restore normal glycemic levels. In type 1 DM and Gestational diabetes a treatment to substitute insulin or insulin analogues is applied. In type 2 DM similar treatment may be applied, or treatment with oral antidiabetics. To determine if the treatment is adequate a test called the glycated hemoglobin, HbA1c test, is carried out. The test provides average blood glucose levels[9] over a period of two to three months. A non-diabetic[10] person has an $HbA1c < 6\%$. In patients with type 1 DM the treatment is correct only if the HbA1c test result stays below the target value.
2) Patient's perspective: the information provided in real-time about blood glucose and HbA1c tests allows the patient to have their disease exhaustively tracked without being hospitalized. Moreover, the telemedicine system is able to detect any anomaly in the patient's treatment. If an anomaly occurs, an alert will be sent to the physician staff. The physician will take the appropriate steps, for example, changing the medication or the dose of a specific drug.

C. DATA DESCRIPTION

For the purpose of checking the performance of our architecture we have used different types of information sources. We have used data extracted from the Web and structured data extracted from a data repository in the data collection phase. Data collected from sensors has been used in the running phase. Specifically, we have experimented with:

1) Web data. The type of unstructured and semi-structured information from the Internet (including social networks) that is continuously gathered is unlimited. However, in this example, we are going to restrict the information to the specific context of two websites: PubMed and PLOS ONE. PubMed is an excellent place to obtain information as it comprises more than

[9]Blood glucose (BG) concentration will vary even in persons with normal pancreatic hormonal function. A normal BG ranges approximately 80-140 mg/dl. The target range for a person with diabetes mellitus is very controversial. It would be very desirable to keep 90% of all BG measurements < 200 mg/dl and that the average BG should be 150 mg/dl or less.

[10]Target values of HbA1c are established by different organisms. They are slightly different. For instance, the American Diabetes Association – ADA– defined in 2005 the target A1c of 7.5%-8% or less whereas the National Institute for Health and Care Excellence of United Kingdom – NICE– published in 2016 the updated guidelines that recommend a target A1c level of 6.5% or lower.

24 million citations of biomedical literature from MEDLINE, life science journals, and online books. PLOS ONE publication is a multidisciplinary Open Access journal. PLOS ONE's broad scope provides a platform to publish primary research, including interdisciplinary and replication studies as well as negative results. It features reports of original research from all disciplines within science and medicine.

2) Structured database. We have used a highly referenced website with structured information ready for data mining using several AI methods: the UCI database. The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms.
3) Sensor data. Currently, there are not only a lot of inexpensive devices for collecting data from the human body, but smartphones can also gather this information.

As has been shown in Figure 1, there are other possible sources of information (the nodes of our architecture) but for reasons of space constraints of this paper we have restricted our sources to these three very significant nodes as they represent a fully operational set.

D. WEB RULES EXTRACTION

The steps to obtain the Web rules, previously explained in Section III-A.1, are the following:

- Health sector document selection (PubMed and PLOS ONE docs).
- IR task to obtain the relevant documents for the physician's needs (use of keywords such as HbA1c, blood glucose, race, age, etc.).
- Semantic tagging of the concepts found in the relevant documents retrieved (use of the domain ontology, UMLS in our example).
- Partial parsing and IE task to identify the document entities.
- Ontology mapping (UMLS-WordNet) to identify equivalent or compatible concepts between different resources.
- Extraction of Web rules using predefined patterns with WordNet concepts.

Next, we are going to explain in detail the mentioned steps. The two first steps are the document selection and the IR task (Figure 2).

With respect to the document selection, we have chosen documents related to the health sector (mentioned in Section III-A.1) from PubMed and PLOS ONE URLs. These documents are of a wide variety of themes and topics. Therefore, it is necessary to carry out a filter to select only documents about diabetes. This is achieved by applying the next step, the IR task. As previously mentioned in Section III-A.1, the objective of this task is to retrieve the documents relevant to the user needs. The user requirements are introduced by the experts (the physicians) as keywords in natural

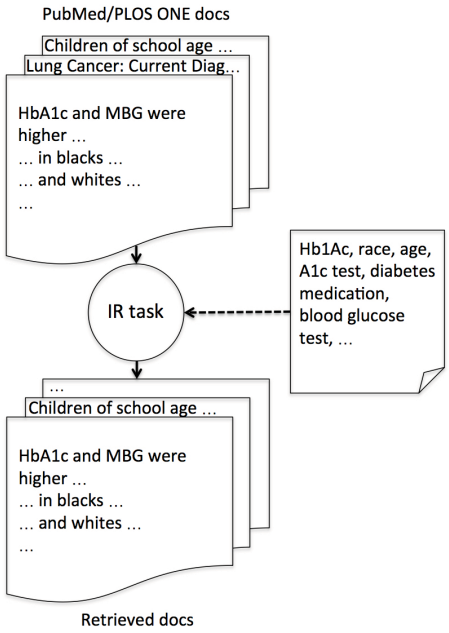

FIGURE 2. Process to obtain Web rules: document selection and IR task.

language related to the main topic: diabetes. In Figure 2 a set of keywords are shown: HbA1c, race, age, A1c test, diabetes medication, etc. The output of this step is the set of relevant documents to the keywords. For instance, in the figure, the document about lung cancer is discarded because is not related to diabetes disease.

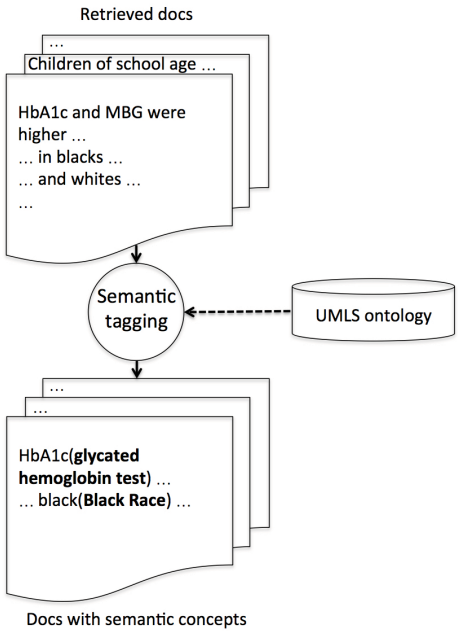

FIGURE 3. Process to obtain Web rules: semantic tagging.

The next step is the semantic tagging of the concept documents (Figure 3). We have used the domain ontology UMLS. In our proposal, one of the main goals is to conceptualize the

different terms and expressions in the texts obtained from the Web. As seen in the previous state-of-the-art section, UMLS is the most frequently used resource by NLP systems in the diverse specialized subdomains in the areas of medicine and health. For the purposes of conceptualization, we consider that UMLS is the best resource because of its wide variety of specialized sources in different subdomains of the medicine and health sectors.

In Figure 3 an example of the semantic tagging can be observed. For example, from the PubMed document with URL code "https://www.ncbi.nlm.nih.gov/pubmed/26783014" (visited on 19th of March, 2018) the following text was extracted: *HbA1c and MBG were higher ($p < 0.0001$) in blacks [10.4% (90.3 mmol/mol), 255 mg/dL] than whites [8.9% (73.9 mmol/mol), 198 mg/dL).*

TABLE 1. Free-text entries mapped to UMLS.

Free-text	Concept CUI	CN	Semantic TUI	STY
Hba1c	C0373638	GLYCATED HEMOGLOBIN TEST	T059	Laboratory Procedure
MBG	C0392201	Blood glucose measurement	T059	Laboratory Procedure
blacks	C0005680	Black race	T098	Population Group
whites	C0007457	Caucasoid Race	T098	Population Group

Table 1 shows the labeling of the UMLS concepts of the previous text. On the one hand, the CUI column uniquely identifies the concept while the CN column shows the name of the concept, and on the other hand, the TUI column uniquely identifies the semantic type while the STY column describes the name of the semantic type associated to the concept.

The output of this step is the semantic tagging with UMLS concepts of the relevant documents. For example, the words HbA1c and black of the relevant document are tagged with the UMLS concepts (CN columns) "*glycated hemoglobin test*" and "*black race*" respectively (Figure 3).

Once the semantic tagging of the concepts has been performed, a set of rules are extracted from the partial parsing of the text in the following step (Figure 4). In this way, each clause that contains any semantic link will be converted into a rule. For example, in the clause "*HbA1c and MBG were higher ($p < 0.0001$) in blacks [10.4% (90.3 mmol/mol), 255 mg/dL] than whites [8.9% (73.9 mmol/mol), 198 mg/dL).*", the parsing will tag the following noun phrases (*np*), verbal phrases (*vp*), and prepositional phrases (*pp*):

np(HbA1c),
np(MBG),
vp(be),
pp(in, *np*(blacks, 10.4%, 90.3 mmol/mol, 255 mg/dL)),
np(whites, 8.9%, 73.9 mmol/mol, 198 mg/dL).

Furthermore, after applying the IE task, a set of entities are extracted. In our example, three types of entities which express measures are identified: (i) percentage (10.4, 8.9), (ii) mmol/mol (90.3, 73.9), and (iii) mg/dL (255, 198).

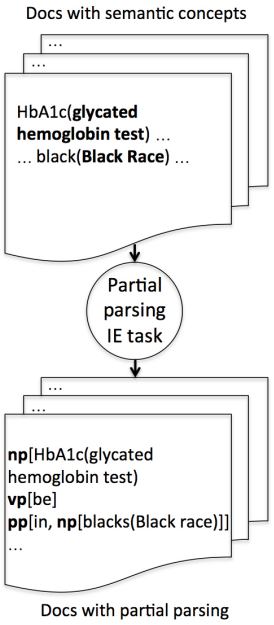

FIGURE 4. Process to obtain Web rules: partial parsing.

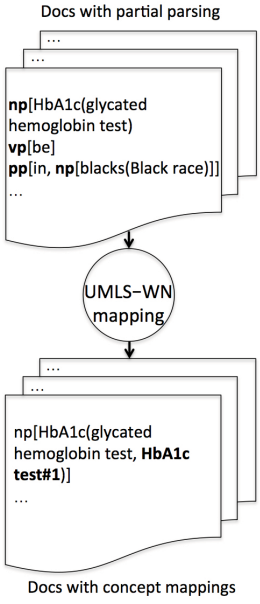

FIGURE 5. Process to obtain Web rules after the ontology mapping between UMLS and WordNet concepts.

In the following step, the ontology mapping between the concepts of the domain ontology (UMLS) and the core ontology (WordNet) is carried out (Figure 5). The STROMA methodology presented in Section III-B is used. For instance, we obtain that the UMLS concept "GLYCATED HEMOGLOBIN TEST" and the WordNet concept "HbA1c test#1" are equivalent. The output of this stage will contain the concepts tagged with the core ontology which will allow the data integration.

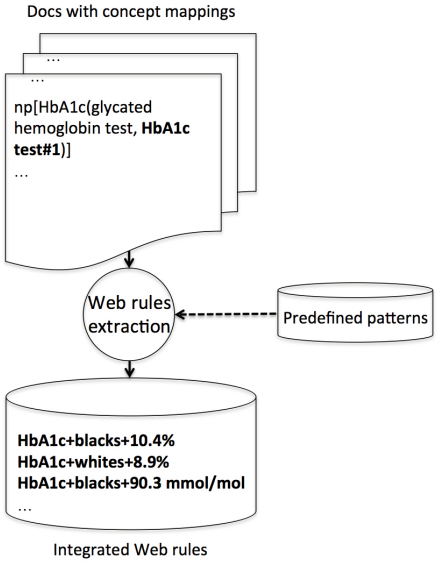

FIGURE 6. Process to obtain Web rules: Web rules extraction using the predefined patterns.

Finally, in the last step, the integrated Web rules that contain relevant information about diabetes are extracted thanks to a series of predefined patterns that use WordNet concepts (Figure 6). In the above example, the following two patterns related to HbA1c and MBG (Mean Blood Glucose) are used:

1) "HbA1c test" concept + "Race" concept + + "measure_percentage" entity ‖ "measure_mmol/mol" entity

2) "Blood glucose test" concept + "Race" concept + + "measure_mg/dL" entity

For example, the first pattern specifies that an "HbA1c test" concept (relative to the HbA1c test) has been detected in the text. It is followed by a "Race" concept (relative to the person's race) and an entity that expresses a measure in percentage or mmol/mol (the value of the HbA1c test result).

After applying the aforementioned patterns, the Web rules shown in Figure 7 are obtained.

HbA1c + blacks + 10.4%
HbA1c + whites + 8.9%
HbA1c + blacks + 90.3 mmol/mol
HbA1c + whites + 73.9 mmol/mol
MBG + blacks + 255 mg/dL
MBG + whites + 198 mg/dL

FIGURE 7. Web rules obtained after data integration. Influence of the patient's race.

These rules obtained from Web documents (Web rules) are very important because they determine that there is a racial variation of HbA1c test result in patients with similar conditions. This result is provided to the specialist physician. With this information the physician can create a personalized treatment for the patient depending on their race with a more accurate analysis of the alarms produced by the telemedicine system. Finally, the Web rules will be later used by physicians and computer experts to improve the DM rules used by the telemedicine system.

Another very illustrative example was extracted from the PLOS ONE document with URL code "http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0152332" (visited on 19th of March, 2018). The following text was found: *Children of school age (6 to 12 years) = target A1c of 8% or less; Adolescents and young adults (between 13−19 years) = target A1c of 7.5% or less.*

In this example the following pattern is used:

3) "measure_age" entity + "HbA1c test" concept + "measure_percentage" entity ‖ "measure_mmol/mol" entity

6 to 12 years + A1c + 8%
13-19 years + A1c + 7.5%

FIGURE 8. Web rules obtained after data integration. Influence of the patient's age.

Consequently, the Web rules shown in Figure 8 are obtained. With this information the physician can create a personalized treatment for the patients depending on their ages. When an alarm is sent to a young patient, the physician will verify his/her exact age and will determine if the A1c value is correct or abnormal.

E. DM RULES EXTRACTION

The steps to extract the DM rules, presented in Section III-A.2, are the following:

- Document selection (UCI Machine Learning Repository).
- Semantic tagging of the concepts found in the database fields (use of the domain ontology, Cyc in our case study).
- Application of AI methods to obtain the DM rules enriched with semantic concepts.
- Ontology mapping (Cyc-WordNet) to identify equivalent or compatible concepts between different resources.

Next, we are going to explain in detail the mentioned steps. The two first steps are the document selection and the semantic tagging (Figure 9).

In terms of the document selection, we have used the structured data extracted from patients with diabetes used in the study developed by Strack *et al.* [113] and has been extracted from the Health Facts database (Cerner Corporation, Kansas City, MO), a national data warehouse that collects comprehensive clinical records across hospitals throughout the United States.

The database contains data systematically collected from participating institutions' electronic medical records and includes encounter data (emergency, outpatient, and inpatient), provider specialty, demographics (age, sex, and race), diagnoses and in-hospital procedures documented by

ICD-9-CM codes, laboratory data, pharmacy data, in-hospital mortality, and hospital characteristics.

The Health Facts data we used was an extract representing 10 years (1999-2008) of clinical care at 130 hospitals and integrated delivery networks throughout the United States. The database consists of 41 tables in a fact-dimension schema and a total of 117 features. The database includes 74,036,643 unique encounters (visits) that correspond to 17,880,231 unique patients and 2,889,571 providers.

The data set was created in two steps. First, encounters of interest were extracted from the database with 55 attributes. This data set is available as Supplementary Material available online,[11] and it is also in the UCI Machine Learning Repository.

Second, preliminary analyses and preprocessing of the data were performed resulting in only these features (attributes) and encounters that could be used in further analyses being retained, in other words, features that contain sufficient information. The full list of the features and their description is provided in [113]. It is important to emphasize the specific attributes related to diabetes: ''glucose serum test result'' (blood glucose level), ''A1c test result'', ''diabetes medications'' and ''change of medications'' (indicates if there was a change in diabetic medications –either dosage or generic name–) among others.

Finally, the information from the mentioned database for ''diabetic'' encounters was extracted. In this way, 101,766 encounters were identified related to diabetic patients. This data was used in our experiments.

The next step is the semantic tagging of the abovementioned ''diabetic DB'' using the domain ontology. We have used the Cyc ontology[12] in our experiments because it stores a large volume of information and has a wealth of links to other ontologies/semantic resources (including health domains).

Specifically, to tag the structured data we have looked for all the attributes of the DB in Cyc. For instance, the attributes ''A1c test result'' and ''Race'' are tagged with the Cyc concepts ''*#HgbA1cBloodTest*'' and ''*#BilogicalSubspecies*'' respectively (Figure 9). The objective of this stage is to identify the semantic concept[13] of each field of the DB in order to subsequently perform the integration of different data sources.

The following step consists of applying Artificial Intelligence methods with the objective to mine structured data to uncover hidden knowledge. We used decision trees to obtain a set of DM rules extracted from the data related to the treatment of patients with diabetes. This approach offers excellent results, as well as the benefit of the easy and efficient visualization of the data. The experiments were performed using Weka software [114]. The algorithm chosen in our study is C4.5 [115]. Tree C4.5 shows, in the form of a tree, the conditions for predicting the behaviour of a specified variable; furthermore, the set of DM rules or conditions are presented in the form of text. In our case study, we reduced the original number of attributes to 18 (the main ones related to diabetes: ''race'', ''age'', ''glucose serum test result'', ''A1c test result'', ''diabetes medications'', ''change of medications'', etc.).

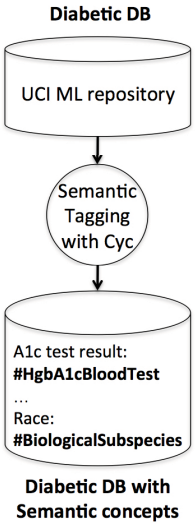

FIGURE 9. Process to obtain DM rules: document selection and semantic tagging.

From the point of view of the physician and the patient, one of the most important things is to analyse whether the patient's treatment is adequate or not. Where it is not adequate, the physician must change the drugs or their doses in order to avoid complications of the disease. This attribute exists in the DB and it is called ''change of medications''. Therefore, our experiments are focused on predicting the value of this variable (which uses the values of *Ch* and *No* to indicate whether the medication should be changed or not respectively) depending on the conditions that occur. A fragment of the obtained DM rules is shown in Figure 10.

```
...
|  max_glu_serum = >200
|  |  A1Cresult = Norm: No (5.0/1.0)
|  |  A1Cresult = >7
|  |  |  number_inpatient <= 0: No (5.0)
|  |  |  number_inpatient > 0: Ch (2.0)
|  |  A1Cresult = >8: Ch (21.0/9.0)
|  |  A1Cresult = None: Ch (1105.0/382.0)
...
|  max_glu_serum = >300: Ch (837.0/174.0)
...
```

FIGURE 10. Excerpt of the DM rules obtained with C4.5 algorithm in ''diabetic database''.

An important fact to note is that the system learns through the patient's individual characteristics (age, weight, sex, etc.), which determine the patient's personalized treatment. Moreover, with the DM rules learned automatically we have already filtered and eliminated many situations that are

[11] http://dx.doi.org/10.1155/2014/781670 (visited on 19th of March, 2018).

[12] http://www.opencyc.org/ (visited on 19th of March, 2018).

[13] As previous mentioned in Section III-A.1 WSD techniques should be applied if it is necessary to obtain an unique semantic concept.

not dangerous for the patient. Finally, these DM rules are supervised by a medical specialist to detect errors or to make improvements to them.

In the DM rules obtained in the experiment, the first factor that influences the change of the patient's medication is the blood glucose test result, ''glucose serum test result''. In the DB this variable has 4 values: (1) *Norm* (normal value), (2) *None* (no glucose measure), (3) >200 y (4) >300.

We focus on the DM rules obtained when the glucose measurement is anomalous (if the value is *Norm* or *None*, no immediate action is taken because there is not any alarm and, initially, no warning is sent to the system). The rules show that if the glucose value is >200 (it is a high value), the A1c test result must be checked. Depending on its value and whether the patient has been previously admitted to the hospital the change of the patient's medication will be decided. Where the medication needs to be changed, an alarm is sent to the system for review by the specialist. Finally, the rules indicate that if the glucose measurement is >300 (a very high value) the medication will be changed. This rule will be reviewed by a specialist who must confirm or modify it.[14]

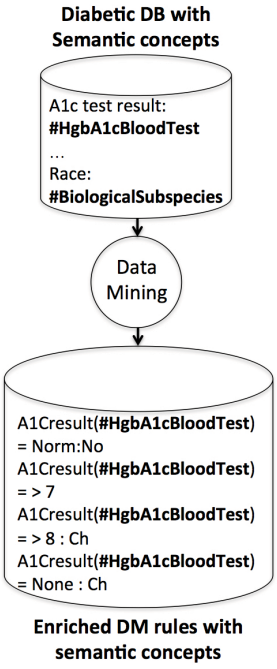

FIGURE 11. Process to obtain DM rules enriched with Cyc concepts after Data Mining.

The output of this step is the DM rules obtained with AI methods (in our experimentation tree C4.5) enriched with semantic concepts. A similar approach to include semantic of rules from C4.5 has been used in [116].

[14]It usually happens that an outlier value in the glucose serum test does not imply a change of medication. Normally, it is necessary to check the latest values of the blood glucose test or to perform the A1c test that provides average blood glucose levels over the past two to three months.

In Figure 11 the enriched DM rules can be observed. For example, in the rules in which the A1cresult attribute of the database appears, its corresponding Cyc semantic concept has been added: A1Cresult(*#HgbA1cBloodTest*) = Norm: No, A1Cresult(*#HgbA1cBloodTest*) = >7, etc.

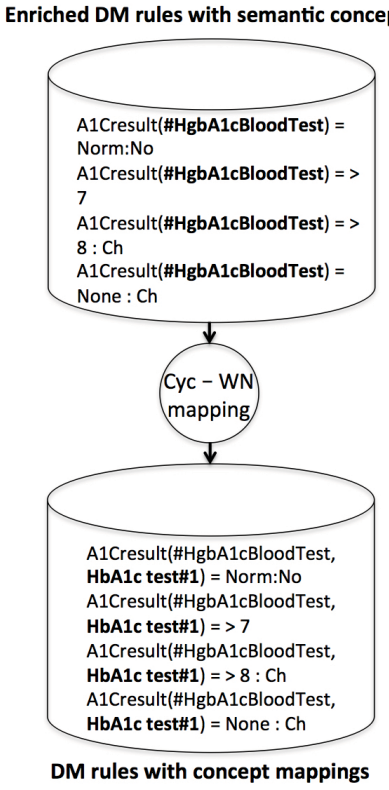

FIGURE 12. Process to obtain DM rules after the ontology mapping between Cyc and WordNet concepts.

In the final step, the ontology mapping between the concepts of the domain ontology (Cyc) and the core ontology (WordNet) is carried out (Figure 12). For instance, we obtain that the Cyc concept ''#HgbA1cBloodTest'' and the WordNet concept ''HbA1c test#1'' are equivalent. This information is included in the rules obtaining the DM rules enriched with the WordNet semantic concepts: A1Cresult(#HgbA1cBloodTest, *HbA1c test#1*) = Norm: No, A1Cresult(#HgbA1cBloodTest, *HbA1c test#1*) = >7, etc. The final result of the DM rules extraction process is a set of enriched DM rules in which the attributes are labeled with their semantic concepts using the core ontology allowing the integration of the different data sources.

These automatically obtained DM rules are very important because they are used by the telemedicine system. The DM rules determine the change of the patient's medication and the sending of the alarm to the physician. We can summarize that the patient's quality of life has considerably improved: (1) the use of sensors (which perform blood glucose and A1c tests) allows their health status to be monitored in real time, without visiting the hospital; (2) if there is no alarm, it is not necessary for a doctor's visit or face-to-face monitoring;

(3) the physician can remotely ''monitor'' the patient's condition in real time and ''adjust'' their medication in a more efficient way. Furthermore, it is important to mention that the clinical costs are also reduced for these mentioned reasons.

F. SENSOR DATA. RUNNING PHASE

The results of this stage for sensor data is the real-time collection of specialized patient data using sensors. This is carried out in the running phase. The information collected from the sensors is sent to the telemedicine system in real-time. If anomalous data is observed/processed, the system will send an alarm to the medical team so that the data is analysed by them.

In our experiments, we were interested in monitoring the main Key Performance Indicators (KPIs) of Type 1 Diabetes patients during their daily physical activities. As seen in the previous state-of-the-art section, considering the large number of sensors that can be monitored to extract biomedical data from patients, we have focused on those that give more information about Type 1 Diabetes. Thus, we will use a sensor to perform the blood glucose measurement control and another one to perform the A1c or HbA1c test at the time intervals specified by the physician. By means of these main tests and other possible scores such as patients' physical effort and sleeping time, the system has enough knowledge to predict abnormal behaviour of this part of the endocrine system in the patients and broadcasts alerts to the medical team.

Moreover, our proposal allows for other sensors that carry out other interesting measurements for this type of patients to be added. For example, cholesterol measurement control, heart rate, and blood pressure levels could be incorporated into the telemedicine system. All this information sent by the sensors is useful if it can be integrated with the remaining information sources. For example, if the sensors send information about heart rate, the DM rules (extracted from the structured data) should contain the ''heart rate'' attribute or equivalent to carry out the data integration.

For the purpose of monitoring the patients, each of them should have different sensors on their bodies that track all these scores and send them to the server, which checks all the scores in order to detect abnormal behaviour. The communication between the different sensors worn by the patients and the server that checks and stores these scores in real-time can be carried out by a dedicated smartphone that uses wifi or 4G to send these scores to the server.

As abovementioned, in the evaluation of our approach we have used the information provided by the sensors related to blood glucose and A1c tests. When the information is received by the telemedicine system, it is checked with the DM rules learned in the data collection phase to decide if there is an alarm (it is necessary a change in the patient's medication).

In Figure 13 a new fragment of the obtained DM rules can be observed. The rules show that if the result of the glucose serum test does not exist ($max\_glu\_serum = None$), the A1c test result must be checked. If A1c test is very high ($=>8$) the patient's medication will be changed depending on his/her age and race. If the patient's age is [20-30] and his/her race is AfricanAmerican and he/she has been previously admitted to hospital (at most one time) the medication must be changed and the alarm is produced by the system. At this moment, the specialists analyse if an exception or anomalous data have been produced. They check the integrated Web rules (Figure 7 and Figure 8) obtained in the Web rules extraction process. As previously mentioned, in Figure 7, a relation between HbA1c (*''HbA1c test'' concept*) and the patient's race (*''Race'' concept*) has been detected establishing that in black people the target value of HbA1c test is 10.4% whereas in white people the target value is 8.9%. Subsequently, there is an exception of the learned DM rule in patients with $race = AfricanAmerican$. Therefore the alarm is discarded and the DM rule is improved with this exception for further decisions. On the contrary, if the A1c is $=>8$ and the patient's age is [10-20] the DM rule says that it is not necessary to change the medication. However, according to extracted Web rules, Figure 8, a relation between HbA1c (*''HbA1c test'' concept*) and the patient's age (*''measure_age'' entity*) has been found. The rules specify that the target A1c of child from 6 to 12 years old could reach 8%. Subsequently, in this case, the medication's change must be carried out. For these reasons, the learned DM rules need to be improved and refined with the rules extracted from the Web (in our experiments extracted from the PubMed and PLOS ONE documents).

```
...
|  max_glu_serum = None
|  |  number_emergency <= 0
...
|  |  |  A1Cresult = >8
|  |  |  |  age = [0-10): No (74.0/11.0)
|  |  |  |  age = [10-20): No (210.0/66.0)
|  |  |  |  age = [20-30)
|  |  |  |  |  race = Caucasian
|  |  |  |  |  |  readmitted = NO: No (43.89/17.0)
|  |  |  |  |  |  readmitted = >30: Ch (18.89/5.89)
|  |  |  |  |  |  readmitted = <30
|  |  |  |  |  |  |  gender = Female: Ch (2.0)
|  |  |  |  |  |  |  gender = Male: No (5.45/2.0)
|  |  |  |  |  |  |  gender = Unknown/Invalid: Ch (0.0)
|  |  |  |  |  race = AfricanAmerican
|  |  |  |  |  |  number_inpatient <= 1: Ch (66.89/27.89)
|  |  |  |  |  |  number_inpatient > 1: No (7.47/2.0)
|  |  |  |  |  race = Other: Ch (6.2/2.2)
|  |  |  |  |  race = Asian: Ch (1.03/0.03)
|  |  |  |  |  race = Hispanic: Ch (5.16/1.16)
|  |  |  |  age = [30-40): Ch (300.0/111.0)
...
```

FIGURE 13. Exceptions of the DM rules obtained with C4.5 algorithm in ''diabetic database''.

We have carried out a second experiment, including new information that will refine and improve the telemedicine system. For this purpose, we have used the specific data obtained from each patient's blood glucose test sensors. The diabetes files of 70 patients from the UCI ML repository have been used and each patient's file has many

FIGURE 14. Daily pre-breakfast blood glucose measurements for patient 02. Predicted values for the next 5 days using LR.

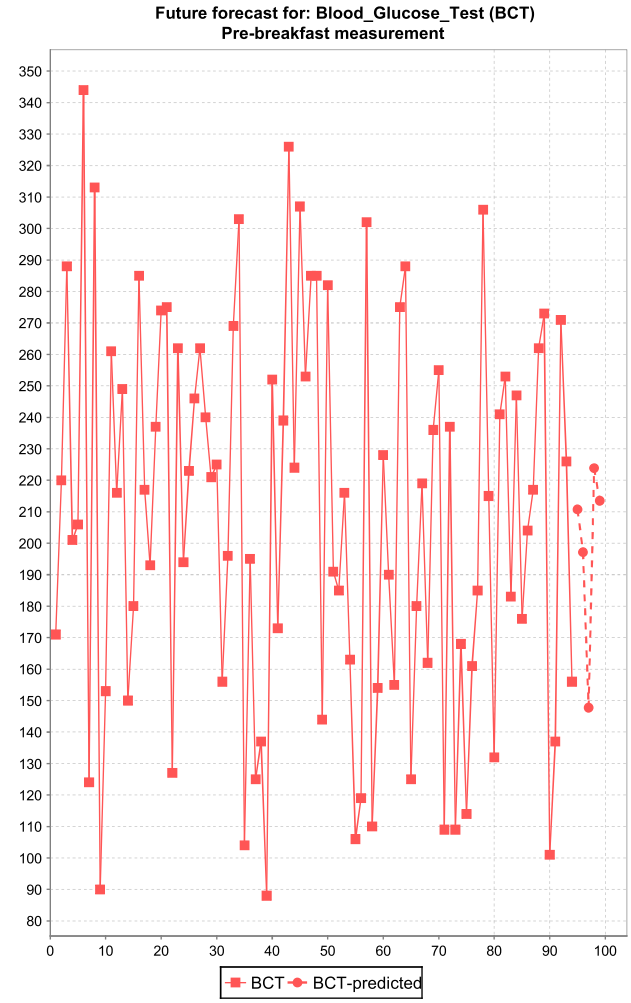

FIGURE 15. Daily pre-breakfast blood glucose measurements for patient 02. Predicted values for the next 5 days using SVM.

records with glucose measurements and other data taken at different intervals. There are four fields per patient record which contain the following information: date, time, descriptive code, and value. The descriptive code field specifies different types of information: Regular/NPH/UltraLente insulin dose, pre-breakfast/pre-lunch/pre-supper blood glucose measurement, hypoglycemic symptoms, typical meal ingestion, etc.

For each patient we have selected the daily blood glucose measurement information. Specifically, the pre-breakfast, pre-lunch, and pre-supper blood glucose measurements have been chosen. The objective is to analyse the blood glucose evolution during the day for each specific patient to facilitate a personalized treatment. Consequently, personalized alarms will trigger when set limits have been surpassed. We use Weka Software for time series analysis to forecast future time steps from the incoming historical data. In the first test, we used the data obtained every day before breakfast for the same patient (in the example we used the data of patient 02 for 94 consecutive days). The predictions of pre-breakfast blood glucose test for the next 5 days using Linear Regression (LR) and Support Vector Machines (SVM) approaches are shown in Figures 14 and 15. The results of the blood glucose test are shown on the ordinate axis and the different measurements (in sequential order) are shown on the abscissa axis.

As observed in Figures 14 and 15 there are significant differences between the predictions using LR and SVM. In the experiment, for day 95, LR predicts a value of 188.6 while SVM predicts a value of 210.7. In our proposal, we have decided to establish a prediction range obtained with the different approaches. This range will indicate a reasonable (non-anomalous) measurement of blood glucose for the following day. For example, taking into account the history of pre-breakfast blood glucose measurements, for patient 02, a range of [188.6-210.7] is established for the measurement of day 95. Similarly, predictions for lunch and supper measures would be carried out. These new rules extracted from daily and personalized information will be included in the

FIGURE 16. Daily pre-breakfast, pre-lunch and pre-supper blood glucose measurements for patient 02. Predicted values for the next 3 days using LR.

FIGURE 17. Daily pre-breakfast, pre-lunch and pre-supper blood glucose measurements for patient 02. Predicted values for the next 3 days using SVM.

telemedicine system with the aim of improving the alarm system.

Correspondingly, in order to improve the predictions of the pre-breakfast measurement, we did another experiment. For this, we also used the pre-lunch and pre-supper measurements. The idea is that the pre-breakfast measurement does not depend exclusively on the measurements obtained in the breakfast of the previous days, but also depends on the measurements obtained in the lunch and supper of the previous days. The predictions of pre-breakfast, pre-lunch and pre-supper blood glucose tests for the next 3 days (9 predictions) are shown in Figures 16 and 17.

In Figures 16 and 17 there are no significant differences between the predictions using LR and SVM. In the experiment, for measure 282 (breakfast on day 95), LR predicts a value of 188.1 while SVM predicts a value of 188.2. In this case, a range of [188.1-188.2] is established for the pre-breakfast measure of day 95. In conclusion, we will incorporate these more precise rules –obtained with the three daily measurements– in the telemedicine system.

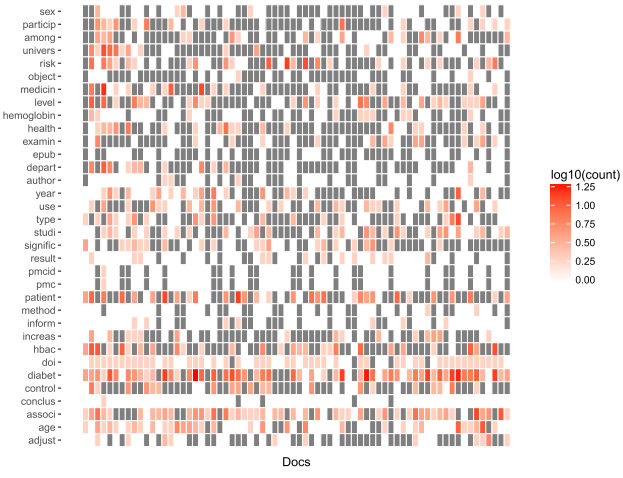

FIGURE 18. Example of word frequencies over PubMed and PLOS ONE documents.

Finally, to complete the experiments, we have carried out an analysis to determine the topics (attributes) of the most used concepts in the documents selected in the web rules

FIGURE 19. Example of visual Text Mining over PubMed and PLOS ONE documents.

extraction process (PubMed and PLOS ONE documents). The main attributes according to their frequency in the documents are represented in Figure 18 and Figure 19. The colour or the letter's size of each attribute represented in both figures indicate their frequency. These graphs corroborate the most outstanding attributes of the documents (*diabetes*, *hbac*, *age*, *sex*, etc.) and that have been expressed in the Web extracted rules.

Thanks to our system, this entire process is developed in real-time where the few seconds the medical team need to make the right decision are crucial to the patient's life. Thus, the main contribution of our system is that it provides the medical team with real-time information for making the right medical decisions to prevent possible deteriorations of Type 1 Diabetes patients.

V. CONCLUSION

In this paper, some of the most novel trends in medicine have been presented. State-of-the-art literature has revealed that traditional methods are not enough, at least on their own, to deal with the current big data situation. The architecture proposed in this paper manages massive data and to carry out efficient experiments according to the complexity of big data scenarios. An ontology-oriented architecture has been defined where a core ontology (WordNet) has been used as a KB enabling data integration of the different heterogeneous sources which use diverse ontologies.

The approach has been applied in the field of personalized medicine (study, diagnosis, and treatment of diseases customized for each patient). AI methods have been used with the objective to mine data in the healthcare sector to uncover knowledge hidden in heterogeneous data sources. A set of learned rules (using Data Mining techniques on structured data, DM rules) and their improvements (applying NLP techniques on data from the Web) are obtained.

A case study in the diabetes scenario has been shown to prove the validity of the proposed model. A telemedicine system that helps the physician provide diabetes treatment and make decisions has been presented. The system allows the physician to improve the DM rules by integrating the information gathered from specialized Web documents. With this new information, the physician can create a personalized treatment for the patient (depending on their specific characteristics) and analyse the alarms produced in the telemedicine system more accurately.

The main novelties presented are the following: (1) an ontology-oriented architecture that uses a central ontology that permits communication between different data sources each with its own ontology; (2) an improvement of the traditional AI systems on diabetes' treatment. The personalized treatment of each patient and the improvement of traditional AI systems have been made possible by including different data sources; and (3) the application of the proposed architecture in the telemedicine system in order to improve its performance.

In a short-term future study, and as a part of an ongoing project, the objective is to improve the performance of another of the V's of Big Data. In particular Velocity, by incorporating a flexible framework for real-time embedded systems. This project provides solutions for a wide range of devices with very heterogeneous capabilities for which it is difficult to predict response times [117]. In future work, we plan to include new data sources such as Social Networks. Furthermore, automating the process for improving DM rules, including rules obtained after applying NLP techniques on Web data (Web rules), must be addressed.

REFERENCES

[1] *Mobile Health Competence Centre*. Accessed: Mar. 19, 2018. [Online]. Available: http://healthybluebits.com/mobile- health-competence-centre/
[2] *New Health App Apple*. Accessed: Mar. 19, 2018. [Online]. Available: https://www.apple.com/ios/whats-new/health/
[3] *New Health App Android*. Accessed: Mar. 19, 2018. [Online]. Available: https://play.google.com/store/search?q= HEALTH&c=apps
[4] *Sensors Healthcare*. Accessed: Mar. 19, 2018. [Online]. Available: https://play.google.com/store/search?q=SENSORHEALTHCARE
[5] J. F. Colom, H. Mora, D. Gil, and M. T. Signes-Pont, "Collaborative building of behavioural models based on Internet of Things," *Comput. Elect. Eng.*, vol. 58, pp. 385–396, Feb. 2017.
[6] D. Gil, A. Ferrández, H. Mora-Mora, and J. Peral, "Internet of Things: A review of surveys based on context aware intelligent services," *Sensors*, vol. 16, no. 7, p. 1069, 2016.
[7] J. Lanza *et al.*, "Managing large amounts of data generated by a smart city Internet of Things deployment," *Int. J. Semantic Web Inf. Syst.*, vol. 12, no. 4, pp. 22–42, 2016.
[8] M. Fazio, A. Celesti, A. Puliafito, and M. Villari, "Big data storage in the cloud for smart environment monitoring," *Procedia Comput. Sci.*, vol. 52, pp. 500–506, Jun. 2015.
[9] S. Aslam, S. ul Islam, A. Khan, M. Ahmed, A. Akhundzada, and M. K. Khan, "Information collection centric techniques for cloud resource management: Taxonomy, analysis and challenges," *J. Netw. Comput. Appl.*, vol. 100, pp. 80–94, Dec. 2017.
[10] R. Nachiappan, B. Javadi, R. N. Calheiros, and K. M. Matawie, "Cloud storage reliability for big data applications: A state of the art survey," *J. Netw. Comput. Appl.*, vol. 97, pp. 35–47, Nov. 2017.
[11] *The Future of Medical Devices*. Accessed: Mar. 19, 2018. [Online]. Available: http://pathfindersoftware.com/
[12] A. Bagula, D. Djenouri, and E. Karbab, "Ubiquitous sensor network management: The least interference beaconing model," in *Proc. IEEE 24th Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2013, pp. 2352–2356.

[13] M. de Buenaga, D. Gachet, M. J. Maña, J. Mata, L. Borrajo, and E. L. Lorenzo, "IPHealth: Plataforma inteligente basada en open, linked y big data para la toma de decisiones y aprendizaje en el ámbito de la salud," *Procesamiento Lenguaje Natural Rev.*, vol. 55, pp. 161–164, Sep. 2015.
[14] D. Kumar, "The personalised medicine: A paradigm of evidence-based medicine," *Annali dell'Istituto superiore di sanità*, vol. 47, no. 1, pp. 31–40, 2011.
[15] G. H. Fernald, E. Capriotti, R. Daneshjou, K. J. Karczewski, and R. B. Altman, "Bioinformatics challenges for personalized medicine," *Bioinformatics*, vol. 27, no. 13, pp. 1741–1748, 2011.
[16] D. Gachet, M. de Buenaga, E. Puertas, and M. T. Villalva, "Big data processing of bio-signal sensors information for self-management of health and diseases," in *Proc. 9th Int. Conf. Innov. Mobile Internet Services Ubiquitous Comput. (IMIS)*, Blumenau, Brazil, L. Barolli, F. Palmieri, H. D. S. Silva, and H. Chen, Eds. Piscataway, NJ, USA: IEEE, Jul. 2015, pp. 330–335.
[17] J. Medina, M. Espinilla, Á. L. García-Fernández, and L. Martínez, "Intelligent multi-dose medication controller for fever: From wearable devices to remote dispensers," *Comput. Elect. Eng.*, vol. 65, pp. 400–412, Jan. 2017.
[18] J. Medina, M. Espinilla, and C. Nugent, "Real-time fuzzy linguistic analysis of anomalies from medical monitoring devices on data streams," in *Proc. 10th EAI Int. Conf. Pervasive Comput. Technol. Healthcare*, 2016, pp. 300–303.
[19] G. Urzaiz, E. Murillo, S. Arjona, R. Hervas, J. Fontecha, and J. Bravo, "An integral medicine taking solution for mild and moderate alzheimer patients," in *Proc. Int. Workshop Ambient Assisted Living*. Berlin, Germany: Springer, 2013, pp. 104–111.
[20] G. K. Savova *et al.*, "Mayo clinical text analysis and knowledge extraction system (cTAKES): Architeture, component evaluation and applications," *J. Amer. Med. Inform. Assoc.*, vol. 17, no. 5, pp. 507–513, 2010.
[21] P. V. Ogren *et al.*, "Constructing evaluation corpora for automated clinical named entity recognition," in *Proc. 12th World Congr. Health (Medical) Inform., Building Sustain. Health Syst. (Medinfo)*, 2007, p. 2325.
[22] R. Pivovarov, A. J. Perotte, E. Grave, J. Angiolillo, C. H. Wiggins, and N. Elhadad, "Learning probabilistic phenotypes from heterogeneous EHR data," *J. Biomed. Inform.*, vol. 58, pp. 156–165, Dec. 2015.
[23] C. Soguero-Ruiz *et al.*, "Predicting colorectal surgical complications using heterogeneous clinical data and kernel methods," *J. Biomed. Inform.*, vol. 61, pp. 87–96, Jun. 2016.
[24] Q. T. Zeng, S. Goryachev, S. Weiss, M. Sordo, S. N. Murphy, and R. Lazarus, "Extracting principal diagnosis, co-morbidity and smoking status for asthma research: Evaluation of a natural language processing system," *BMC Med. Inform. Decis. Making*, vol. 6, no. 1, p. 30, 2006.
[25] J.-W. Fan and C. Friedman, "Semantic reclassification of the UMLS concepts," *Bioinformatics*, vol. 24, no. 17, pp. 1971–1973, 2008.
[26] D. Demner-Fushman, W. W. Chapman, and C. J. McDonald, "What can natural language processing do for clinical decision support?" *J. Biomed. Inform.*, vol. 42, no. 5, pp. 760–772, 2009.
[27] R. M. Terol, P. Martínez-Barco, and M. Palomar, "A knowledge based method for the medical question answering problem," *Comput. Biol. Med.*, vol. 37, no. 10, pp. 1511–1521, 2007.
[28] S. T. Wu *et al.*, "Unified medical language system term occurrences in clinical notes: A large-scale corpus analysis," *J. Amer. Med. Inform. Assoc.*, vol. 19, no. e1, pp. e149–e156, 2012.
[29] R. Xu, M. A. Musen, and N. H. Shah, "A comprehensive analysis of five million UMLS metathesaurus terms using eighteen million MEDLINE citations," in *Proc. AMIA Annu. Symp.*, 2010, p. 907.
[30] M. Jiang *et al.*, "A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries," *J. Amer. Med. Inform. Assoc.*, vol. 18, no. 5, pp. 601–606, 2011.
[31] R. J. Carroll *et al.*, "Portability of an algorithm to identify rheumatoid arthritis in electronic health records," *J. Amer. Med. Inform. Assoc.*, vol. 19, no. e1, pp. e162–e169, 2012.
[32] I. De Falco, "Differential evolution for automatic rule extraction from medical databases," *Appl. Soft Comput.*, vol. 13, no. 2, pp. 1265–1283, 2013.
[33] A. R. Bhavsar and H. A. Arolkar, "Multidimensional association rule based data mining technique for cattle health monitoring using wireless sensor network," in *Proc. Int. Conf. Comput. Sustain. Global Develop. (INDIACom)*, 2014, pp. 810–814.
[34] S. Batra, H. J. Parashar, S. Sachdeva, and P. Mehndiratta, "Applying data mining techniques to standardized electronic health records for decision support," in *Proc. 6th Int. Conf. Contemporary Comput. (IC3)*, 2013, pp. 510–515.
[35] M. Grossglauser and H. Saner, "Data-driven healthcare: From patterns to actions," *Eur. J. Preventive Cardiol.*, vol. 21, no. 2, pp. 14–17, 2014.
[36] H. C. Koh and G. Tan, "Data mining applications in healthcare," *J. Healthcare Inf. Manage.*, vol. 19, no. 2, pp. 64–72, 2011.
[37] E. AbuKhousa and P. Campbell, "Predictive data mining to support clinical decisions: An overview of heart disease prediction systems," in *Proc. Int. Conf. Innov. Inf. Technol. (IIT)*, Mar. 2012, pp. 267–272.
[38] D. Tomar and S. Agarwal, "A survey on data mining approaches for healthcare," *Int. J. Bio-Sci. Bio-Technol.*, vol. 5, no. 5, pp. 241–266, 2013.
[39] S. Soni and O. P. Vyas, "Building weighted associative classifiers using maximum likelihood estimation to improve prediction accuracy in health care data mining," *J. Inf. Knowl. Manage.*, vol. 12, no. 1, pp. 1350008-1–1350008-14, 2013.
[40] A. Subasini, N. F. Abubacker, and C. Rekha, "Analysis of classifier to improve medical diagnosis for breast cancer detection using data mining techniques," *Int. J. Adv. Netw. Appl.*, vol. 5, no. 6, pp. 2117–2122, 2014.
[41] T. Zhu, Y. Ning, A. Li, and X. Xu, "Using decision tree to predict mental health status based on Web behavior," in *Proc. 3rd Symp. Web Soc.*, 2011, pp. 27–31.
[42] B. Castellani and J. Castellani, "Data mining: Qualitative analysis with health informatics data," *Qualitative Health Res.*, vol. 13, no. 7, pp. 1005–1018, 2003.
[43] F. Ibrahim, M. Ati, and B. El Nagar, "Diagnosis of diabetes mellitus in an E-health environment based on artificial neural network," *J. Next Gener. Inf. Technol.*, vol. 4, no. 5, pp. 125–132, 2013.
[44] T. Joachims, *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*, vol. 186. Norwell, MA, USA: Kluwer, 2002.
[45] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*. Amsterdam, The Netherlands: Elsevier, 2008.
[46] D. Chung, K. C. Lee, and S. C. Seong, "General Bayesian network approach to health informatics prediction: Emphasis on performance comparison," *Procedia-Social Behav. Sci.*, vol. 81, pp. 465–468, Jun. 2013.
[47] B. A. Thakkar, M. I. Hasan, and M. A. Desai, "Health care decision support system for swine flu prediction using Naïve Bayes classifier," in *Proc. Int. Conf. Adv. Recent Technol. Commun. Comput. (ARTCom)*, 2010, pp. 101–105.
[48] T. Balasubramanian and R. Umarani, "Clustering as a data mining technique in health hazards of high levels of fluoride in potable water," *Int. J. Adv. Comput. Sci. Appl.*, vol. 3, no. 2, pp. 166–171, 2012.
[49] B. Pogorelc and M. Gams, "Recognition of patterns of health problems and falls in the elderly using data mining," in *Proc. 17th Iberoamerican Congr. Prog. Pattern Recognit., Image Anal., Comput. Vis., Appl. (CIARP)*, 2012, pp. 463–471.
[50] P. A. Bath, "Data mining in health and medical information," *Annu. Rev. Inf. Sci. Technol.*, vol. 38, no. 1, pp. 331–369, 2004.
[51] S.-M. Zhou, M. A. Rahman, M. Atkinson, and S. Brophy, "Mining textual data from primary healthcare records: Automatic identification of patient phenotype cohorts," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2014, pp. 3621–3627.
[52] X. Li, X. Liu, Z. Zhang, Y. Xia, and S. Qian, "Design of health eating system based on Web data mining," in *Proc. WASE Int. Conf. Inf. Eng. (ICIE)*, Aug. 2010, pp. 346–349.
[53] R. Nagavelli and C. V. G. Rao, "Degree of disease possibility (DDP): A mining based statistical measuring approach for disease prediction in health care data mining," in *Proc. Int. Conf. Recent Adv. Innov. Eng.*, May 2014, pp. 1–6.
[54] Y. Demchenko, Z. Zhao, P. Grosso, A. Wibisono, and C. de Laat, "Addressing big data challenges for scientific data infrastructure," in *Proc. IEEE 4th Int. Conf. Cloud Comput. Technol. Sci. (CloudCom)*, Dec. 2012, pp. 614–617.
[55] T. B. Murdoch and A. S. Detsky, "The inevitable application of big data to health care," *J. Amer. Med. Assoc.*, vol. 309, no. 13, pp. 1351–1352, 2013.
[56] H. Banaee, M. U. Ahmed, and A. Loutfi, "Data mining for wearable sensors in health monitoring systems: A review of recent trends and challenges," *Sensors*, vol. 13, no. 12, pp. 17472–17500, 2013.
[57] G. W. Hunter *et al.*, "Smart sensor systems for human health breath monitoring applications," *J. Breath Res.*, vol. 5, no. 3, p. 037111, 2011.
[58] E. Kantoch, "Technical verification of applying wearable physiological sensors in ubiquitous health monitoring," in *Proc. Comput. Cardiol. (CinC)*, Sep. 2013, pp. 269–272.

[59] W. D. Bae, S. Alkobaisi, S. Narayanappa, and C. C. Liu, "A real-time health monitoring system for evaluating environmental exposures," *J. Softw.*, vol. 8, no. 4, pp. 791–801, 2013.
[60] P. Velardi, G. Stilo, A. E. Tozzi, and F. Gesualdo, "Twitter mining for fine-grained syndromic surveillance," *Artif. Intell. Med.*, vol. 61, no. 3, pp. 153–163, 2014.
[61] A. Nikfarjam and G. H. Gonzalez, "Pattern mining for extraction of mentions of adverse drug reactions from user comments," in *Proc. AMIA Annu. Symp.*, 2011, pp. 1019–1026.
[62] K. L. P. Nguyen and N. Ashish, "Large scale, complex processing of health data with MapReduce," *J. Inf. Knowl. Manage.*, vol. 13, no. 1, pp. 1450009-1–1450009-6, 2014.
[63] H. Q. Yu, X. Zhao, X. Zhen, F. Dong, E. Liu, and G. Clapworthy, "Healthcare-event driven semantic knowledge extraction with hybrid data repository," in *Proc. 4th Ed. Int. Conf. Innov. Comput. Technol. (INTECH)*, Aug. 2014, pp. 13–18.
[64] C.-H. Lin, L.-C. Huang, S.-C. T. Chou, C.-H. Liu, H.-F. Cheng, and I.-J. Chiang, "Temporal event tracing on big healthcare data analytics," in *Proc. IEEE Int. Congr. Big Data*, Jun./Jul. 2014, pp. 281–287.
[65] E. A. Mohammed, B. H. Far, and C. Naugler, "Applications of the MapReduce programming framework to clinical big data analysis: Current landscape and future trends," *BioData Mining*, vol. 7, p. 22, Oct. 2014.
[66] D. Dubois, W. Liu, J. Ma, and H. Prade, "The basic principles of uncertain information fusion. an organised review of merging rules in different representation frameworks," *Inf. Fusion*, vol. 32, pp. 12–39, Nov. 2016.
[67] W. Sujansky, "Heterogeneous database integration in biomedicine," *J. Biomed. Inform.*, vol. 34, no. 4, pp. 285–298, 2001.
[68] A. Bronselaer, D. V. Britsom, and G. D. Tre, "Propagation of data fusion," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 5, pp. 1330–1342, May 2015.
[69] S. Konieczny and R. P. Pérez, "Logic based merging," *J. Philos. Logic*, vol. 40, no. 2, pp. 239–270, 2011.
[70] A. Bilke, J. Bleiholder, F. Naumann, C. Böhm, K. Draba, and M. Weis, "Automatic data fusion with HumMer," in *Proc. 31st Int. Conf. Very Large Data Bases (VLDB)*, 2005, pp. 1251–1254.
[71] F. Naumann, A. Bilke, J. Bleiholder, and M. Weis, "Data fusion in three steps: Resolving inconsistencies at schema-, tuple-, and value-level," *Bull. Tech. Committee Data Eng.*, vol. 29, no. 2, pp. 21–31, 2006.
[72] L. Cholvy and S. Moral, "Merging databases: Problems and examples," *Int. J. Intell. Syst.*, vol. 16, no. 10, pp. 1193–1221, 2001.
[73] C. Cobelli *et al.*, "Pilot studies of wearable outpatient artificial pancreas in type 1 diabetes," *Diabetes Care-Alexandria*, vol. 35, no. 9, pp. e65–e67, 2012.
[74] D. Machado, T. Paiva, I. Dutra, V. S. Costa, and P. Brandão, "Managing diabetes: Pattern discovery and counselling supported by user data in a mobile platform," in *Proc. IEEE Symp. Comput. Commun. (ISCC)*, Jul. 2017, pp. 296–299.
[75] V. Oncescu, M. Mancuso, and D. Erickson, "Cholesterol testing on a smartphone," *Lab Chip*, vol. 14, no. 4, pp. 759–763, 2014.
[76] W.-J. Yi, W. Jia, and J. Saniie, "Mobile sensor data collector using Android smartphone," in *Proc. IEEE 55th Int. Midwest Symp. Circuits Syst. (MWSCAS)*, Aug. 2012, pp. 956–959.
[77] G. Fortino and V. Giampà, "PPG-based methods for non invasive and continuous blood pressure measurement: An overview and development issues in body sensor networks," in *Proc. IEEE Int. Workshop Med. Meas. Appl. (MeMeA)*, Apr./May 2010, pp. 10–13.
[78] G. Sannino and G. De Pietro, "A mobile system for real-time context-aware monitoring of patients' health and fainting," *Int. J. Data Mining Bioinf.*, vol. 10, no. 4, pp. 407–423, 2014.
[79] J.-L. Reyes-Ortiz, L. Oneto, A. Samà, X. Parra, and D. Anguita, "Transition-aware human activity recognition using smartphones," *Neurocomputing*, vol. 171, pp. 754–767, Jan. 2016.
[80] A. Doan, A. Halevy, and Z. Ives, *Principles of Data Integration*. Amsterdam, The Netherlands: Elsevier, 2012.
[81] B. Golshan, A. Halevy, G. Mihaila, and W.-C. Tan, "Data integration: After the teenage years," in *Proc. 36th ACM SIGMOD-SIGACT-SIGAI Symp. Princ. Database Syst.*, 2017, pp. 101–106.
[82] M. Lenzerini, "Data integration: A theoretical perspective," in *Proc. 21st ACM SIGMOD-SIGACT-SIGART Symp. Princ. Database Syst.*, 2002, pp. 233–246.
[83] S. Kopparty and B. Rossman, "The homomorphism domination exponent," *Eur. J. Combinatorics*, vol. 32, no. 7, pp. 1097–1114, 2011.
[84] T. Jayram, P. G. Kolaitis, and E. Vee, "The containment problem for real conjunctive queries with inequalities," in *Proc. 25th ACM SIGMOD-SIGACT-SIGART Symp. Princ. Database Syst.*, 2006, pp. 80–89.
[85] F. Afrati, C. Li, and P. Mitra, "Rewriting queries using views in the presence of arithmetic comparisons," *Theor. Comput. Sci.*, vol. 368, nos. 1–2, pp. 88–123, 2006.
[86] J. Goldstein and P.-Å. Larson, "Optimizing queries using materialized views: A practical, scalable solution," *ACM SIGMOD Rec.*, vol. 30, no. 2, pp. 331–342, 2001.
[87] M. Zaharioudakis, R. Cochrane, G. Lapis, H. Pirahesh, and M. Urata, "Answering complex SQL queries using automatic summary tables," *ACM SIGMOD Rec.*, vol. 29, no. 2, pp. 105–116, 2000.
[88] K. Runapongsa, J. M. Patel, H. V. Jagadish, Y. Chen, and S. Al-Khalifa, "The Michigan benchmark: Towards XML query performance diagnostics," *Inf. Syst.*, vol. 31, no. 2, pp. 73–97, 2006.
[89] L. Segoufin and V. Vianu, "Validating streaming XML documents," in *Proc. 21st ACM SIGMOD-SIGACT-SIGART Symp. Princ. Database Syst.*, 2002, pp. 53–64.
[90] I. Tatarinov, S. D. Viglas, K. Beyer, J. Shanmugasundaram, E. Shekita, and C. Zhang, "Storing and querying ordered XML using a relational database system," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2002, pp. 204–215.
[91] S. Abiteboul, O. Benjelloun, B. Cautis, I. Manolescu, T. Milo, and N. Preda, "Lazy query evaluation for active XML," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2004, pp. 227–238.
[92] S. Abiteboul, A. Bonifati, G. Cobéna, I. Manolescu, and T. Milo, "Dynamic XML documents with distribution and replication," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2003, pp. 527–538.
[93] A. Gangemi, R. Navigli, and P. Velardi, "The OntoWordNet project: Extension and axiomatization of conceptual relations in WordNet," in *Proc. OTM Confederated Int. Conf. Move Meaningful Internet Syst.* Berlin, Germany: Springer, 2003, pp. 820–838.
[94] D. Suciu, D. Olteanu, C. Ré, and C. Koch, "Probabilistic databases," *Synthesis Lectures Data Manage.*, vol. 3, no. 2, pp. 1–180, 2011.
[95] X. Dong, A. Y. Halevy, and C. Yu, "Data integration with uncertainty," in *Proc. 33rd Int. Conf. Very Large Data Bases*, 2007, pp. 687–698.
[96] A. Gal, "Uncertain schema matching," *Synthesis Lectures Data Manage.*, vol. 3, no. 1, pp. 1–97, 2011.
[97] M. Lu, D. Agrawal, B. T. Dai, and A. K. H. Tung, "Schema-as-you-go: On probabilistic tagging and querying of wide tables," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2011, pp. 181–192.
[98] B. He and K. C.-C. Chang, "Statistical schema matching across Web query interfaces," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2003, pp. 217–228.
[99] M. Magnani, N. Rizopoulos, P. M. Brien, and D. Montesi, "Schema integration based on uncertain semantic mappings," in *Conceptual Modeling—ER*. Berlin, Germany: Springer, 2005, pp. 31–46.
[100] S. Kumar, N. Baliyan, and S. Sukalikar, "Ontology cohesion and coupling metrics," *Int. J. Semantic Web Inf. Syst.*, vol. 13, no. 4, pp. 1–26, 2017.
[101] D. Berardi, D. Calvanese, and G. De Giacomo, "Reasoning on UML class diagrams," *Artif. Intell.*, vol. 168, nos. 1–2, pp. 70–118, 2005.
[102] Y. Arens, C. Y. Chee, C.-N. Hsu, and C. A. Knoblock, "Retrieving and integrating data from multiple information sources," *Int. J. Intell. Cooperat. Inf. Syst.*, vol. 2, no. 2, pp. 127–158, 1993.
[103] E. Bytyçi, B. Sejdiu, A. Avdiu, and L. Ahmedi, "SEMDPA: A semantic Web crossroad architecture for WSNs in the Internet of Things," *Int. J. Semantic Web Inf. Syst.*, vol. 13, no. 3, pp. 1–21, 2017.
[104] N. F. Noy and M. A. Musen, "Algorithm and tool for automated ontology merging and alignment," in *Proc. 17th Nat. Conf. Artif. Intell. (AAAI)*, 2000, pp. 1–6.
[105] M. Ganzha, M. Paprzycki, W. Pawłowski, P. Szmeja, and K. Wasielewska, "Semantic interoperability in the Internet of Things: An overview from the INTER-IoT perspective," *J. Netw. Comput. Appl.*, vol. 81, pp. 111–124, Mar. 2017.
[106] M. Stevenson and Y. Wilks, "Word sense disambiguation," in *The Oxford Handbook of Computational Linguistics*. London, U.K.: Oxford Univ. Press, 2003, pp. 249–265.
[107] R. Navigli, "Word sense disambiguation: A survey," *ACM Comput. Surv.*, vol. 41, no. 2, 2009, Art. no. 10.
[108] A. Olivé, "The universal ontology: A vision for conceptual modeling and the semantic Web (invited paper)," in *Conceptual Modeling* (Lecture Notes in Computer Science), vol. 10650. Berlin, Germany: Springer, 2017, pp. 1–17.
[109] C. Fellbaum, "WordNet," in *Theory and Applications of Ontology: Computer Applications*. Berlin, Germany: Springer, 2010, pp. 231–243.
[110] C. Matuszek *et al.*, "Searching for common sense: Populating Cyc from the Web," in *Proc. AAAI*, 2005, pp. 1430–1435.

[111] P. Shvaiko and J. Euzenat, "Ontology matching: State of the art and future challenges," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 1, pp. 158–176, Jan. 2013.
[112] P. Arnold and E. Rahm, "Enriching ontology mappings with semantic relations," *Data Knowl. Eng.*, vol. 93, pp. 1–18, Sep. 2014.
[113] B. Strack *et al.*, "Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records," *BioMed Res. Int.*, vol. 2014, Apr. 2014, Art. no. 781670.
[114] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *ACM SIGKDD Explorations Newslett.*, vol. 11, no. 1, pp. 10–18, 2009.
[115] J. R. Quinlan, *C4. 5: Programs for Machine Learning*. Amsterdam, The Netherlands: Elsevier, 2014.
[116] M. Espinilla, J. Medina, Á.-L. García-Fernández, S. Campaña, and J. Londoño, "Fuzzy intelligent system for patients with preeclampsia in wearable devices," *Mobile Inf. Syst.*, vol. 2017, Oct. 2017, Art. no. 7838464.
[117] H. M. Mora, D. Gil, J. F. C. López, and M. T. S. Pont, "Flexible framework for real-time embedded systems based on mobile cloud computing paradigm," *Mobile Inf. Syst.*, vol. 2015, Jun. 2015, Art. no. 652462.

JESÚS PERAL received the Ph.D. degree in computer science from the University of Alicante in 2001. He is currently an Assistant Professor with the Department of Software and Computing Systems, University of Alicante. His main research topics include natural language processing, information extraction, information retrieval, question answering, data warehouses, and business intelligence applications. He has participated in numerous national and international projects, agreements with private companies and public organizations related to his research topics. He has published many papers (over 40 papers) in journals and conferences related to his research interests.

ANTONIO FERRÁNDEZ received the Ph.D. degree in computer science from the University of Alicante in 1998. He is currently an Assistant Professor with the Department of Software and Computing Systems, University of Alicante. His research topics include information extraction, information retrieval, question answering, natural language processing, ellipsis, and anaphora resolution. He has participated in numerous national and international projects, agreements with private companies and public organizations related to his research topics. He has participated in many conferences and most of his work has been published in international journals and conferences, with over 70 published papers.

DAVID GIL is currently an Assistant Professor with the Department of Computing Technology and Data Processing, University of Alicante. His main research topics include artificial intelligence applications, data mining, open data, big data, and decision support system in medical and cognitive sciences. He has participated in numerous national and international projects, agreements with private companies and public organizations related to his research topics. He has participated in many conferences and most of his work has been published in international journals and conferences, with over 50 published papers.

RAFAEL MUÑOZ-TEROL received the master's and Ph.D. degrees in computer science from the University of Alicante in 2002 and 2009, respectively. Since 2002, he has been a Lecturer with the Lucentia Research Group, Department of Software and Computing Systems, University of Alicante. He is currently involved in the area of big data and key performance indicators. His research interests include big data analytics, predictive data mining, and business intelligence systems.

HIGINIO MORA received the B.S. degree in computer science engineering, the B.S. degree in business studies, and the Ph.D. degree in computer science from the University of Alicante, Spain, in 1996, 1997, and 2003, respectively. Since 2002, he has been a Faculty Member with the Computer Technology and Computation Department, University of Alicante, where he is currently an Associate Professor and also a Researcher at the Specialized Processors Architecture Laboratory. His areas of research interests include computer modeling, computer architectures, high performance computing, embedded systems, Internet of Things, and cloud computing paradigm. He has participated in many conferences and most of his work has been published in international journals and conferences, with over 60 published papers.